Winter 2015

Using large SNP datasets to understand the genetic mechanisms of complex traits in Arabidopsis thaliana

Elisabeth S. Harrison
Purdue University

Graduate School Form 30
(Updated 11/20/2014)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Elisabeth Svedin Harrison

Entitled
USING LARGE SNP DATASETS TO UNDERSTAND THE GENETIC MECHANISMS OF COMPLEX TRAITS IN ARABIDOPSIS THALIANA

For the degree of Master of Science

Is approved by the final examining committee:

Brian Dilkes

Clint Chapple

Rebecca Doerge

Michael Gribskov

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Brian Dilkes

Approved by Major Professor(s):

Approved by: Christine Hrycyna 02/06/2015

Head of the Department Graduate Program Date

USING LARGE SNP DATASETS TO UNDERSTAND THE GENETIC MECHANISMS OF COMPLEX TRAITS IN *ARABIDOPSIS THALIANA*

A Thesis

Submitted to the Faculty

of

Purdue University

by

Elisabeth S. Harrison

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

May 2015

Purdue University

West Lafayette, Indiana

Dedicated to my mom

ACKNOWLEDGEMENTS

I would like to acknowledge the help of Tena Graham, who helped me with the flow cytometry data for the tetraploid study. I was also want to acknowledge her help with caring for my plants during the last four years.

I would also like to acknowledge Dr. Stephen Weller for his help with the glufosinate study. He scored the damage of the plant and helped with the glufosinate treatments.

Lasly, I want to acknowledge my husband, Aaron, and my daughter, Kisty, for their support and love. Aaron has supported me during this difficult journey, and has pushed me when I could not go any further. Kisty has been very patient during this journey too.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

ABSTRACT

Harrison, Elisabeth Svedin. M.S., Purdue University, May 2015. Using Large SNP Datasets to Understand the Genetic Mechanisms of Complex Traits in *Arabidopsis thaliana*. Major Professor: Brian Dilkes.

Arabidopsis thaliana, as a model species, has been widely genotyped and sequenced. Many studies have been done to understand the kinship and population structure of the species. This data and information is beneficial for understanding the genetic mechanisms of complex traits. In this thesis, we first used genotyped data for 5,967 accessions to study the occurrence of tetraploidy in the species. We found that tetraploidy is a transient character state, and the species is a diploid species. Secondly, we used 211K and 1.6M SNPs for 440 accessions to run genome-wide analyses (GWA) for four traits: glufosinate tolerance, hybrid incompatibilities, seed size, and secondary metabolites. We used two different statistical methods, EMMAX and MLMM, to calculate the associations between SNP and phenotype. Putative gene lists for each trait from each statistical model are available for the general public to find candidate genes that are involved in these traits.

CHAPTER 1. TETRAPLOIDY IS A TRANSIENT CHARACTER STATE IN *ARABIDOPSIS THALIANA*

1.1 Introduction

Polyploidy, the condition of having three or more complete chromosome sets, has had a major role in plant evolution and speciation (Otto and Whitton 2000, Cui et al. 2006, Wood et al. 2009). Two major types of polyploids have been distinguished: allopolyploids, organisms that have more than two copies of distinct hybridized genomes, and autopolyploids, organisms that have more than two copies of the same genome (Otto and Whitton 2000, Pignatta et al. 2010). In nature, allotetraploids are typically distinguishable from the diploid parents and many are recognized as a different species. Autotetraploids, on the other hand, are neither readily distinguishable from diploid progenitors nor always recognized as different species than the progenitors. In fact, many species are identified with diploid and tetraploid populations (Soltis and Soltis 2000). For these reasons, it has been hypothesized that allopolyploidy is more widespread as a speciation mechanism than autotetraploidy. Recent research has refuted this notion, and now autotetraploidy has been recognized as playing a major role in plant evolution, diversification, and adaptation (Soltis and Soltis 2000, Soltis et al. 2010).

Though autopolyploidization plays a major role in plant evolution, no clear and consistent fitness advantages are described (Comai 2005, Soltis et al. 2010). Through the years many advantages and disadvantages of autopolyploidization have been

hypothesized (Ohno 1970, Doyle 1986, Li et al. 1996, Lynch and Force 2000, Comai 2005, Freeling and Thomas 2006, Saleh et al. 2008, Soltis et al. 2010, Birchler et al. 2010). Any advantage of polyploidy could contribute to persistence of autopolyploids, and any disadvantage could contribute to the extinction of autopolyploids.

Disadvantages that autopolyploids have to overcome initially are decreased reproductive abilities because of minority cytotype exclusion and the production of aneuploid offspring (Levin 1975, Doyle 1986). Minority cytotype exclusion selects against the cytotype that is the minority in a given environment, and puts downward pressures on the likely survival of nascent tetraploid subpopulations. If the tetraploid does not self-pollinate then most of the crossing partners will be diploids, resulting in triploid offspring and disappearance of the tetraploid from the population. Also, postzygotic barriers inhibit the diploid-tetraploid hybridization and the majority of seeds abort (Dilkes et al. 2008). Autotetraploids that self-pollinate still have a minority cytotype disadvantage they have a smaller population and also because their seed set may be smaller than the diploid species (Henry et al. 2005, Chao et al. 2013).

Another disadvantage of polyploidization is the production of aneuploid swarms. Upon polyploidization, the new genome experiences losses of genes and whole genomes and can result in the production of aneuploid offspring (Henry et al. 2006, 2009). The production of aneuploid offspring decreases reproduction because a high percentage of aneuploid offspring are infertile. Aneuploid offspring are evident in *A. thaliana* tetraploids, *Secale cereale* tetraploids, and *Zea mays* tetraploids (Randolph 1935, Müntzing 1951, Henry et al. 2006). Henry et al. (2006) found that more than 25% of the offspring of tetraploids were aneuploid (Henry et al. 2006); therefore, the reproduction

rate of *a. thaliana* tetraploids is decreased by 25%, decreasing the likelihood of tetraploids outcompeting the diploid parental population in the same environment.

Persistence of an autotetraploid population requires drift or selective advantage overcoming these disadvantages (Levin 1975, Rodriguez 1996, Petit et al. 1999). For example, Chao et al (2013) demonstrated that *A. thaliana* tetraploids have a higher salinity tolerance than diploids (Chao et al. 2013). Hypothetically, an *A. thaliana* tetraploid will have a selective advantage over the diploid progenitor in a soil with high salinity. The diploid will have a lower reproductive rate than the tetraploid and essentially remove the minority cytotype and aneuploid production disadvantages (Chao et al. 2013). Further differentiation and diversification of the autotetraploid population from the diploid population, together with the established postzygotic isolation between diploids and tetraploids, would contribute to eventual polyploidy speciation events.

Arabidopsis thaliana has been prominently used as a model system to study the genomic and genetic consequences of polyploidy (Weiss and Maluszynska 2000, Comai et al. 2000, Henry et al. 2005, Yu et al. 2009). Some of these studies interpreted results assuming that tetraploid *A. thaliana* populations are successful in the wild. A few tetraploid accessions are known, and laboratory-induced tetraploids are stable and can reproduce, but *A. thaliana* is a predominantly diploid species (Yu et al. 2009). Four naturally occurring autotetraploids have been reported to date: Wa-1 (Schmuths et al. 2004, Henry et al. 2005), M3385S (Henry et al. 2005), Bla-5 (Bomblies et al. 2007, Chao et al. 2013), and Ciste-2 (Chao et al. 2013). Though these have been reported as being collected from the wild it is unclear if these autotetraploids come from established (i.e. adapted) populations or if they are rare off-type tetraploids among the many collections

from diploid populations. If they were only rare off-type cytotypes than interpreting studies using these tetraploids may be misleading and not indicate how autotetraploids become established in the wild.

Recently, studies on the genetic variation and population structure in *A. thaliana* demonstrate that its population structure is strong between accessions that are spatially close together (Bergelson et al. 1998, Platt et al. 2010, Bomblies et al. 2010). Outcrossing only occurs ~1-5% of the time, and gene flow between populations is limited (Bergelson et al. 1998, Platt et al. 2010). Platt et al. (2010) found that populations that inhabit the same geographic area were genetically similar and shared haplotypes. They also found that only a small distance (~1 km) was needed to break down the haplotype groups. Geographic distance is sufficient to allow for genetic divergence between populations (Platt et al. 2010).

In this study we investigated the success, defined as established populations in the wild, of tetraploids of *A. thaliana* by testing whether tetraploidy is a persistent or ephemeral character state. We used flow cytometric analysis of nuclear DNA content and genotypic data to compare the four known tetraploids—Wa-1, M3385S, Bla-5, and Ciste-2—to their most closely related accession(s). Firstly, if the tetraploids were collected from persistent populations they should be genetically distinct from collected diploid accessions. Furthermore, if the three tetraploids are subpopulations of a single tetraploid population distributed across Europe they should be more genetically similar to each other than to any other diploid. Alternatively, if they come from persistent independent tetraploid populations then the accessions among ~6000 genotyped accessions most genetically similar to each tetraploid will also be tetraploid. If, however, the tetraploids

are rare off-types they will be genetically similar to their diploid progenitors and the only tetraploid among the genetically similar accessions.

1.2 Methods and Materials

1.2.1 Plant material and genotypes

Plant material was ordered from the ABRC at Ohio State University. The genotype data for each accession is from publically available SNP data from the University of Chicago. 139 SNPs were generated using the Sequenom MassArray system at Sequenom (Platt et al. 2010). 214,553 SNPs were generated using the Affymetrix produced (Affymetrix, Santa Clara, CA) AtSNP Tile Array at the U. Chicago service core and the SNP calls were performed using the Oligo package with modifications (Horton et al. 2012).

1.2.2 Flow cytometry

Flow cytometric analysis of nuclear DNA was used to determine the ploidy level of the accessions. Leaf tissue (several fresh new leaves) or seed (20 μl) was used to isolate nuclei from each accession. Leaf tissue: The leaf tissue was placed in a petri dish with 1 ml of ice-cold chopping buffer (15 mM HEPES, 1 mM EDTA, 80 mM KCl, 20 mM NaCl, 300 mM sucrose, 0.20% triton-X, 0.5 mM spermine, 0.1% beta-mercaptoethanol, stored in 4° C). The leaf tissue was finely chopped with a carbon steel razor blade (VWR #9) for approximately 1 minute, or until the buffer was very green. More buffer was added if needed in order to keep the leaves moist and covered during chopping. Seeds: The seeds were placed in individual wells of a 96-well plate. To each well, 400 ml of chopping buffer was added the mixture ground the seeds with a Zymo

squisher (Zymo Research Corporation, Irvine, CA) or a plastic pestle until the buffer turned a milky white.

Once the tissue was pulverized, 2-4 layers of cheesecloth cut into ~1 cm segments were used to strain the supernatant from the tissue or seeds. The supernatant was transferred to a microcentrifuge tube and placed on ice. Nuclei and other cellular debris were sedimented by centrifugation for 7 min at 500g. The supernatant was discarded without disturbing the pellet by aspirating the liquid by pipette and the pellet was resuspended in 400 ml staining solution (40 μl of 1 mg/ml propidium iodide added to 1 ml chopping buffer). The samples were again centrifuged at 500g for 7 min, the supernatant was discarded and the sample resuspended in another 400 μl of staining buffer. Prepared in this manner, the nuclei are stable when on ice and in the dark. Samples were always incubated 1-2 hours to allow the dye and DNA to reach equilibrium. Samples were run either on the Beckman Coulter Quanta SC or Beckman Coulter FC500 machines (Beckman Coulter, Brea, CA). All the ecotypes that were subjected to flow cytometric analysis and their ploidy levels are presented in Table 1.1.

1.2.3 Determining genetic similarities

To compare the genetic similarity between two accessions we calculated the percentage of SNPs that differ between them. The comparison between the two Wa-1 accessions and Me-0 was made using 214,553 SNPs. The other comparisons were made using the 139 SNPs.

1.2.4 Hierarchical clustering trees

A relationship tree using 5,967 accession was generated using Distance Matrix Computation and Hierarchical Clustering in R: A language and environment for

Table 1.1. Ecotypes and their ploidy level. Flow cytometry was done during this study as indicated by the date, or was published in a previous study as indicated by the reference. Schmuths et al., 2005 and Henry et al., 2005 performed flow cytometry. Bomblies et al., 2007 performed multiple crosses and either detected interploidy lethality (4x) and followed with flow cytometry or did not which we regard as evidence for diploidy.

Ecotype	Accession #	Country	Ploidy	Flow Cytometry
Alc-0	CS1656	Spain	2X	Schmuths et al. 2004
Belmonte-4-94	CS76095	Italy	2X	3/12/13
Bla-1	CS6616	Spain	2X	8/1/10
Bla-10	CS6622	Spain	2X	8/1/10
Bla-11	CS6623	Spain	2X	11/1/10
Bla-2	CS6617	Spain	2X	8/1/10
Bla-3	CS6618	Spain	2X	8/1/10
Bla-5	**CS6620**	**Spain**	**4X**	**8/1/10**
Bla-6	CS6621	Spain	2X	8/1/10
Chi-1	CS6665	Russia	2X	2/1/11
Chi-2	CS6666	Russia	2X	2/1/11
Ciste-1		Italy	2X	8/1/10
Ciste-2		**Italy**	**4X**	**8/1/10**
Cit-0	CS1080	France	2X	Henry et al. 2005
Co-4	N1091	Portugal	2X	Bomblies et al. 2007
Eil-0	N1133	Germany	2X	Bomblies et al. 2007
Er-0	CS1142	Germany	2X	Henry et al. 2005
Fjä1-5	CS76132	Sweden	2X	2/1/11
H55	CS923	Czech Republic	2X	8/1/10
Ha-0	CS1218	Germany	2X	Henry et al. 2005
Hau-0	CS1220	Denmark	2X	Henry et al. 2005
Hl-0	CS1228	Germany	2X	Henry et al. 2005
Ka-0	CS28375	Austria	2X	3/12/13
Kn-0	CS6762	Lithuania	2X	Schmuths et al. 2004
Koch-1	CS22823	Ukraine	2X	8/1/10
Koch-3	CS22825	Ukraine	2X	2/1/11
Koch-7	CS22829	Ukraine	2X	2/1/11
Koch-8	CS28852	Ukraine	2X	2/1/11
Koch-9	CS28853	Ukraine	2X	2/1/11
Ll-0	CS1338	Spain	2X	Henry et al. 2005
Ll-1	CS1340	Spain	2X	Henry et al. 2005
Ll-2	CS6783	Spain	2X	Schmuths et al. 2004
M3385S	**CS6183**	**Sweden**	**4X**	**Henry et al. 2005; 8/1/10**
M7323S	CS6184	Sweden	2X	8/1/10
M7884S	CS6185	Sweden	2X	8/1/10

Table 1.1 Continued

M7943S	**CS6186**	**Sweden**	**4X**	**8/1/10**
Me-0	**CS1364**	**Germany**	**4X**	**8/1/10**
Ms-0	CS6797	Russia	2X	Henry et al. 2005
Old-1	CS28583	Germany	2X	3/12/13
Old-2	CS6821	Germany	2X	3/12/13
Oy-0	CS6824	Norway	2X	Henry et al. 2005
Oy-1	CS1643	Norway	2X	Henry et al. 2005
Pf-0	CS28601	Germany	2X	3/12/13
PHW-1	CS28602	Italy	2X	3/12/13
Pla-0	N14569	Spain	2X	Henry et al. 2005
Pla-1	N1461	Spain	2X	Bomblies et al. 2007
Pla-2	N1463	Spain	2X	Bomblies et al. 2007
Rome-1	CS22524	Italy	2X	3/12/13
Se-0	CS22646	Spain	2X	Henry et al. 2005
Sf-1	CS1512	Spain	2X	Henry et al. 2005
Ste-3	CS76232	USA	2X	3/12/13
Stw-0	N1539	Russia	2X	2/1/11
T510	CS76238	Sweden	2X	3/12/13
Ting-1	CS22549	Sweden	2X	8/1/10
Tiv-1	CS22525	Italy	2X	3/12/13
Ts-1	CS1552	Spain	2X	Henry et al. 2005
Ts-5	CS6871	Spain	2X	Henry et al. 2005
Wa-1	**CS6885, CS22644**	**Poland**	**4X**	**Henry et al. 2005; 8/1/10**
16.2	Bomblies Collection	Llagostera, Spain	2X	7/19/10
22.1	Bomblies Collection	Lloret de Mar, Spain	2X	7/19/10
22.4	Bomblies Collection	Lloret de Mar, Spain	2X	7/19/10
1.-5	Bomblies Collection	La Montgoda, Spain	2X	7/19/10
10.1-1	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
10.1-2	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
13-1	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
13-2	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
15-1	Bomblies Collection	Llagostera, Spain	2X	7/19/10
16-3	Bomblies Collection	Llagostera, Spain	2X	7/19/10
16-4	Bomblies Collection	Llagostera, Spain	2X	7/19/10
16.1-1	Bomblies Collection	Llagostera, Spain	2X	7/19/10
16.1-2	Bomblies Collection	Llagostera, Spain	2X	7/19/10
16.1-3	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-10	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-10-2	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-11	Bomblies Collection	Llagostera, Spain	2X	7/19/10

Table 1.1 Continued

17-13-1	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-13-2	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-3	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-4	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-4-2	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-5	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-6	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-6-2	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-8	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-9	Bomblies Collection	Llagostera, Spain	2X	7/19/10
17-9-2	Bomblies Collection	Llagostera, Spain	2X	7/19/10
18-1-1	Bomblies Collection	Cassà de la Selva, Spain	2X	7/19/10
18-1-2	Bomblies Collection	Cassà de la Selva, Spain	2X	7/19/10
19-1-1	Bomblies Collection	Llambilles, Spain	2X	7/19/10
19-1-2	Bomblies Collection	Llambilles, Spain	2X	7/19/10
19-3-1	Bomblies Collection	Llambilles, Spain	2X	7/19/10
19-3-2	Bomblies Collection	Llambilles, Spain	2X	7/19/10
19-5	Bomblies Collection	Llambilles, Spain	2X	7/19/10
2.-1	Bomblies Collection	Between Pola Giverola & Salionç, Spain	2X	7/19/10
2.-2	Bomblies Collection	Between Pola Giverola & Salionç, Spain	2X	7/19/10
20-1-1	Bomblies Collection	Lloret de Mar, Spain	2X	7/19/10
20-1-2	Bomblies Collection	Lloret de Mar, Spain	2X	7/19/10
20-2-1	Bomblies Collection	Lloret de Mar, Spain	2X	7/19/10
20-2-2	Bomblies Collection	Lloret de Mar, Spain	2X	7/19/10
21-1-1	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
21-1-2	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
21-2-1	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
21-2-2	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
21.1-2	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
22.6-1	Bomblies Collection	Lloret de Mar, Spain	2X	7/19/10
22.6-2	Bomblies Collection	Lloret de Mar, Spain	2X	7/19/10
3.-3	Bomblies Collection	Platja d'Aro, Spain	2X	7/19/10
4.-1	Bomblies Collection	Platja d'Aro, Spain	2X	7/19/10
4.-2	Bomblies Collection	Platja d'Aro, Spain	2X	7/19/10
4.1-2	Bomblies Collection	Platja d'Aro, Spain	2X	7/19/10
4.2-2	Bomblies Collection	Platja d'Aro, Spain	2X	7/19/10
5.-3	Bomblies Collection	Platja d'Aro, Spain	2X	7/19/10
5.-4	Bomblies Collection	Platja d'Aro, Spain	2X	7/19/10

Table 1.1 Continued

5.4-2	Bomblies Collection	Platja d'Aro, Spain	2X	7/19/10
6.-1	Bomblies Collection	Platja d'Aro, Spain	2X	7/19/10
6.1-2	Bomblies Collection	Platja d'Aro, Spain	2X	7/19/10
7.-2	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
7.1-1	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
7.1-2	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10
7.1-3	Bomblies Collection	Tossa de Mar, Spain	2X	7/19/10

statistical computing (R Core Team 2013). The distance matrix used the manhattan method for determining the genetic distance between two individuals. The trees were generated in R using plot command for Hierarchical Clustering. Figures 1.1, 1.3, 1.5, 1.7, and 1.11 show the full tree and the subclades of the tetraploids, Wa-1, M3385S, Bla-5, and Ciste-2, respectively. A sample of the R code used to generate the tree can be found in the Appendix A.

1.2.5 Neighbor-joining trees

Generation of a relationship tree containing 5,967 accessions was done using Dnadist and Neighbor from the phylip-3.69 package (Felsenstein 2004). Dnadist was used to calculate the distance matrix using all 5,967 accessions. Jukes-Cantor was used as the model for nucleotide substitution to calculate the distance matrix. Neighbor was then used to generate a tree representing the relationships between the 5,967 accessions. Figures 1.2, 1.4, 1.6, 1.8, and 1.12 show the full tree and the subclades of the Wa-1, M3385S, Bla-5, and Ciste-2, respectively, as generated by neighbor-joining.

1.3 Results

1.3.1 Wa-1, M3385S, Bla-5, and Ciste-2 are independently derived tetraploids

Three of the previously identified tetraploids, Wa-1, M3385S, and Bla-5 have been genotyped (Platt et al. 2010, Horton et al. 2012). Ciste-2 was not genotyped, but it was sequenced (Cao et al. 2011). To determine the relationships of the four tetraploids, we generated a matrix of the relatedness of 5,967 accessions using their genotypes at 139 SNPs from data previously described (Platt et al. 2010, Anastasio et al. 2011), with the addition of Ciste-1 and Ciste-2 using the sequenced data to extract the base pairs associated with the SNPs (Cao et al. 2011). We generated two trees from these data using

hierarchical clustering and neighbor-joining. Despite differences between these trees, the three tetraploids were not closely related and do not appear to descend from a single tetraploidy event (Figures 1.1 & 1.2). Rather, all three tetraploids belong in different clades within the species-wide genetic diversity.

We also tested the genetic similarities between the four tetraploids using the 139 SNPs. The four tetraploids are not genetically similar, which supports the clustering of the trees (Table 1.2). Therefore, there was no evidence to support the hypothesis of a single tetraploidy event that resulted in the establishment of these four tetraploid populations.

1.3.2 Wa-1 is a unique tetraploid with unclear provenance

The sparse sampling of diploids for flow cytometric determination of nuclear DNA contents (Schmuths et al. 2004, Henry et al. 2005) may have failed to detect additional tetraploid accessions related to the described tetraploid accessions. If there were tetraploid clades we would expect that each known tetraploid is derived from an established tetraploid subpopulation and genetic neighbors would also be tetraploid. Both trees show Wa-1 being closely related to Me-0 (Figures 1.3 & 1.4). Previous work by Anastasio et al. (2011) indicated that Wa-1 and Me-0 were stock duplications, which both can be tracked back to original collections by Kranz (Anastasio et al. 2011). To better characterize the relatedness of Me-0 and Wa-1 we used 214,553 SNPs present on a AtSNPtile array (Horton et al. 2012) to determine the degree of divergence. Pair-wise comparisons between Me-0 and two different accessions of Wa-2 found that Me-0 had more SNP genotypes in common with CS22644 than CS6885 (Table 1.3). The number of

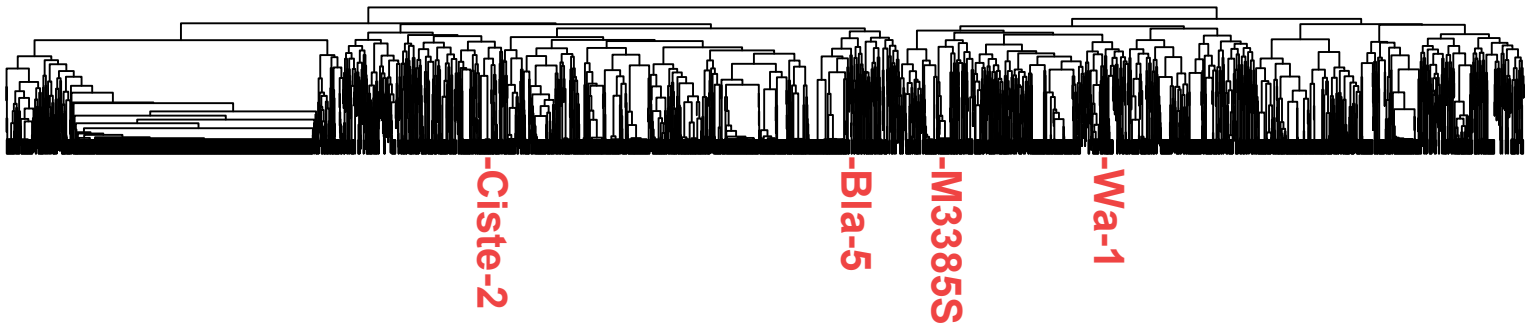

Figure 1.1. Hierarchical clustering demonstrates independent derivation of the tetraploid accessions among 5,967 ecotypes of *A. thaliana*. The tree was generated using 139 SNPs. The manhattan method was used to calculate distances between individuals. The positions of the known tetraploid accessions are indicated.

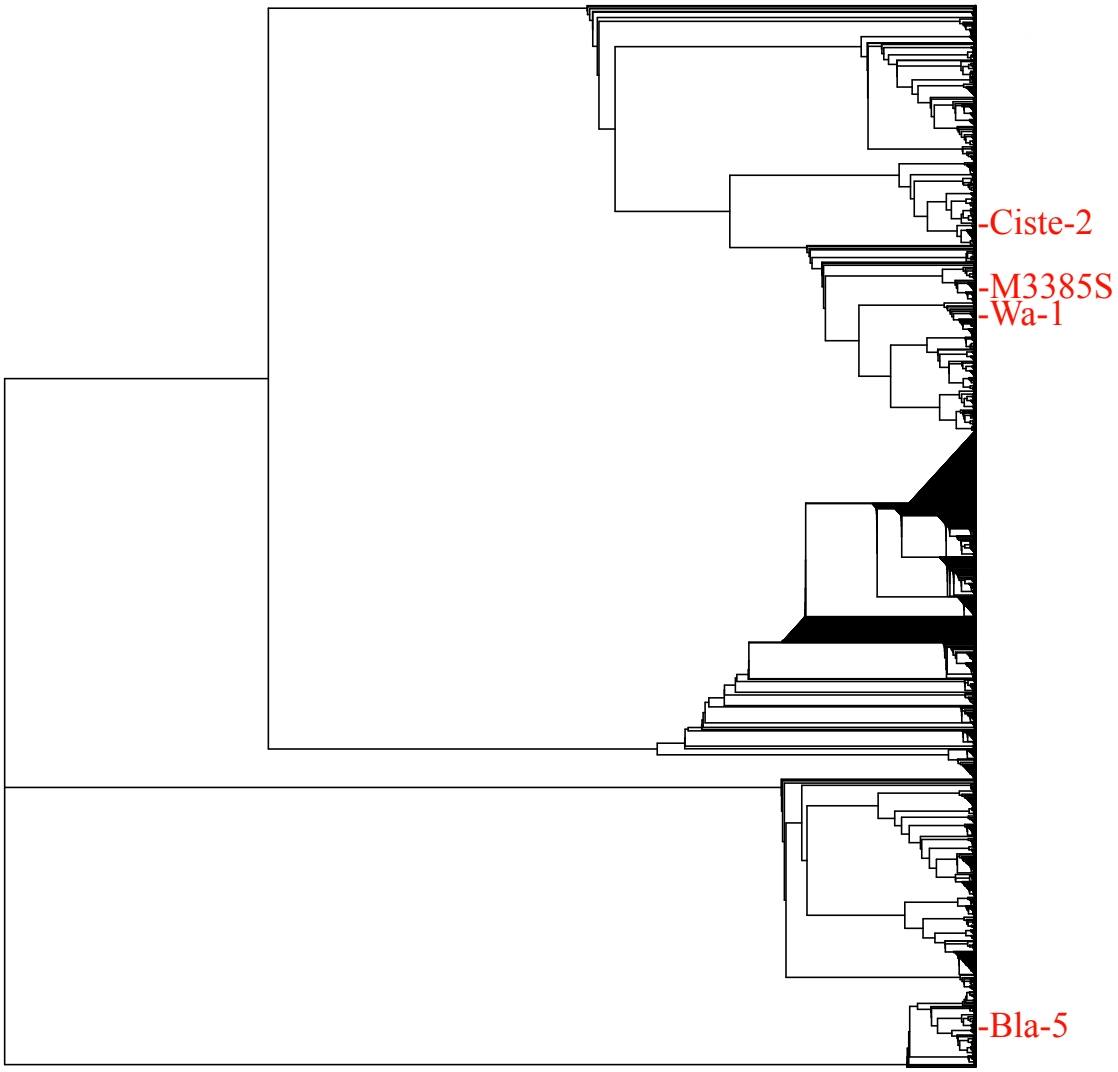

Figure 1.2. Neighbor joining relationship tree demonstrates independent derivation of the tetraploid accessions among 5,967 ecotypes of *A. thaliana*. The tree was generated using 139 SNPs. The Jukes-Cantor method was used to calculate distances between individuals. The positions of the known tetraploid accessions are indicated.

Table 1.2. The frequency of polymorphisms between the tetraploid accessions based on 139 SNPs.

	Ciste-2	**M3385S**	**Bla-5**	**Wa-1**	**Wa-1**
Ciste-2	0	0.403	0.518	0.482	0.475
M3385S		0	0.518	0.496	0.489
Bla-5			0	0.511	0.504
Wa-1				0	0.007
Wa-1					0

Table 1.3. The frequency of polymorphisms between accessions based on 214,553 SNPs.

	Wa-1 (CS6885)	**Wa-1 (CS22644)**	**Me-0**	**Kn-0**
Wa-1 (CS6885)	0	0.048	0.048	0.276
Wa-1 (CS22644)		0	0.053	0.277
Me-0			0	0.277
Kn-0				0

differences between the two Wa-1 accessions and the number of differences between Me-0 and Wa-1 (CS6885) are not significantly different ($\chi^2 = 0.962$, p-value = 0.32). These differences were most likely the result of technical error in the microarray hybridizations and did not provide any evidence favoring genetic divergence of Me-0 from Wa-1. Therefore, these results suggested Me-0 and Wa-1 were the same accession, representing a mislabeled stock duplication. This is consistent with the previous study (Anastasio et al. 2011).

We used flow cytometric analyses to determine the ploidy level of the accessions most closely related to Wa-1. Of the nine other accessions clustered to Wa-1 as determined by hierarchical clustering, only Me-0, the duplicate accession of Wa-1, was tetraploid (Figure 1.3). Additional testing of accessions similar to Wa-1 in the neighbor-joining tree (Figure 1.4) showed that Wa-1 and Me-0 were the only tetraploids and all others were diploids. The flow cytometric results for all accessions that we measured or were measured in other studies (Schmuths et al. 2004, Henry et al. 2005, Bomblies et al. 2007) are available in Table 1.1.

1.3.3 M3385S is a unique tetraploid from Sweden

We looked at M3385S, which is a tetraploid accession collected in Sweden by Napp-Zinn and contributed to the stock center by May and Somerville. The clustering tree and the neighbor-joining tree are remarkably different from each other (Figures 1.5 & 1.6, respectively). Both trees, however, showed that M3385S was most closely related to M7943S, which shares the same provenance as M3385S. We used flow cytometry to determine the ploidy level of the available accessions that were related to M3385S (Table 1.1). Based on the flow cytometric data, M7943S was also tetraploid. Only two

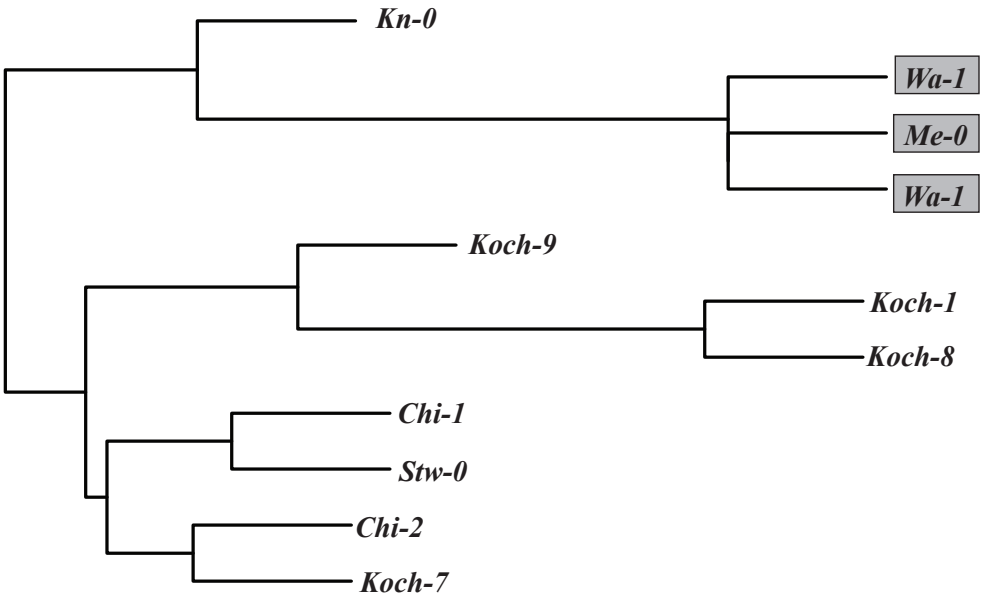

Figure 1.3. Overlay of ploidy and genetic distance for Wa-1 and genetically similar accessions. This subclade is derived from a larger tree consisting of 5,967 accessions based on 139 SNPs. The tree was generated in R using Hierarchical Clustering. Ploidy was determined by flow cytometric analysis of isolated nuclei for all accessions in bold italic. Untested accessions are in normal text, confirmed diploids are in bold italic and tetraploids are boxed.

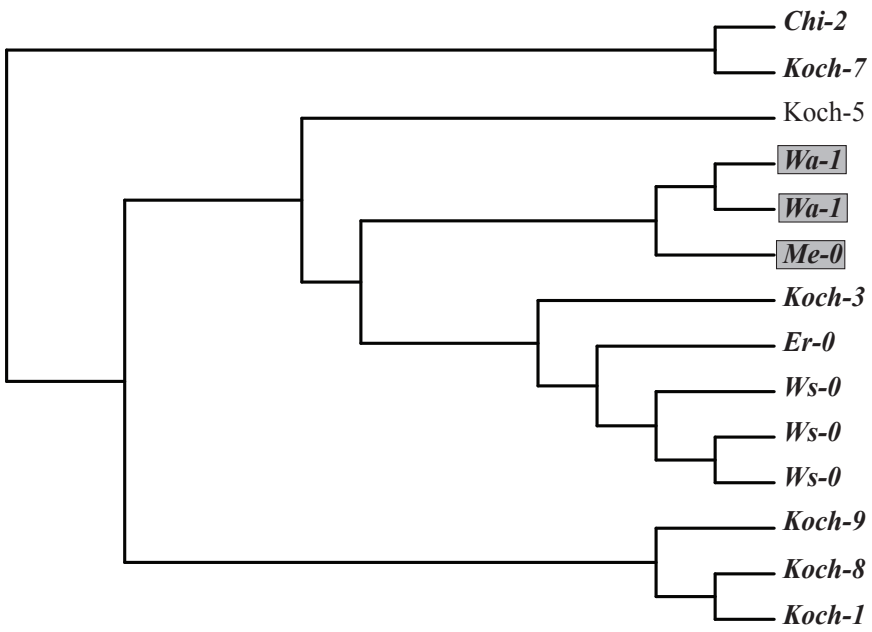

Figure 1.4. Overlay of ploidy and genetic distance for Wa-1 and genetically similar accessions. This subclade is derived from a larger tree consisting of 5964 accessions based on 139 SNPs. The distance matrix was calculated using the jukes-cantor method and the tree was calculated using neighbor joining. Ploidy was determined by flow cytometric analysis of isolated nuclei for all accessions in bold italic. Untested accessions are in normal text, confirmed diploids are in bold italic and tetraploids are boxed.

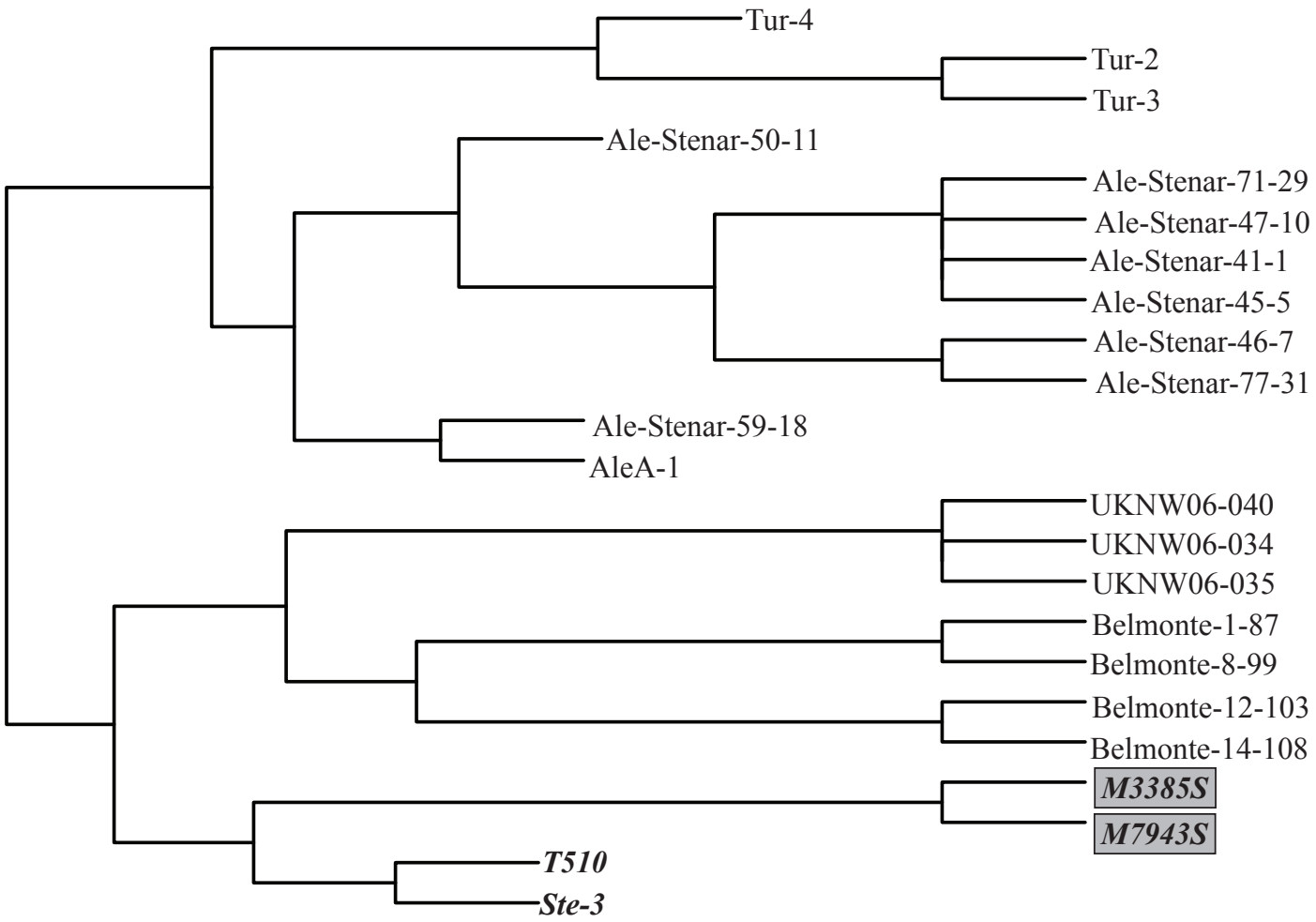

Figure 1.5. Overlay of ploidy and genetic distance for M3385S and genetically similar accessions. This subclade is derived from a larger tree consisting of 5964 accessions based on 139 SNPs. The tree was generated in R using Hierarchical Clustering. Ploidy was determined by flow cytometric analysis of isolated nuclei for all accessions in bold italic. Untested accessions are in normal text, confirmed diploids are in bold italic and tetraploids are boxed.

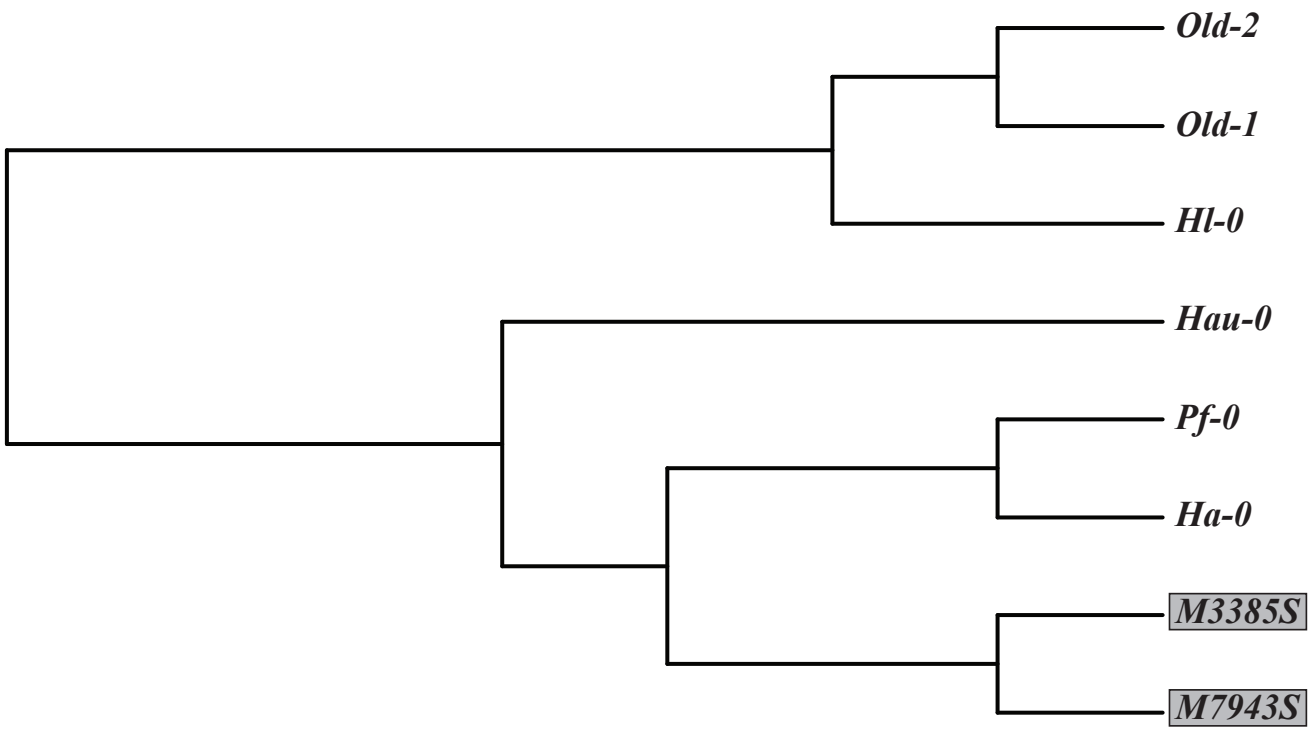

Figure 1.6. Overlay of ploidy and genetic distance for M3385S and genetically similar accessions. This subclade is derived from a larger tree consisting of 5964 accessions based on 139 SNPs. The distance matrix was calculated using the jukes-cantor method and the tree was calculated using neighbor joining. Ploidy was determined by flow cytometric analysis of isolated nuclei for all accessions in bold italic. Untested accessions are in normal text, confirmed diploids are in bold italic and tetraploids are boxed.

accessions, T510 and Ste-3, from the clustering tree were available for testing, and both were diploids. All accessions in the neighboring-joining subclade other than M3385S and M79435S were diploids.

To determine if M3385S and M7943S were two separate populations or the same population, we calculated the genetic similarities between the two accessions using 139 genoytped SNPs. The two accessions were genetically identical (Table 1.4), suggesting that they have had no time to diverge or that they represent duplications of collection or post-collection growouts. Thus, this tetraploid also provided no evidence for tetraploid persistence.

We also calculated the genetic differences for all accessions within the clade. The tetraploid had diverged sufficiently from the most closely related diploid accessions, so much so that no congenic diploid parental ecotype is obvious. This suggests that sampling is not sufficient in Sweden to detect the diploid parent population. While the current data are suggestive of a single tetaploidy event, resampling of the collection site could definitively rule out M3385S/M7943S as an established tetraploid population. Unfortunately, detailed sample location has not been retained with the material.

1.3.4 Bla-5 is identical to a diploid population also found in Blanes, Spain

Bla-5 is a previously detected tetraploid (Bomblies et al. 2007, Chao et al. 2013) and we confirmed this by flow cytometric analysis of nuclear DNA content. The clustering tree in Figure 1.7 shows that Bla-5 was most closely related to Bla-3. The neighbor-joining tree was similar to the clustering tree (Figure 1.8). Comparing the 139 SNPs, Bla-3 and Bla-5 are genetically identical (Table 1.5). If Bla-5 were collected from

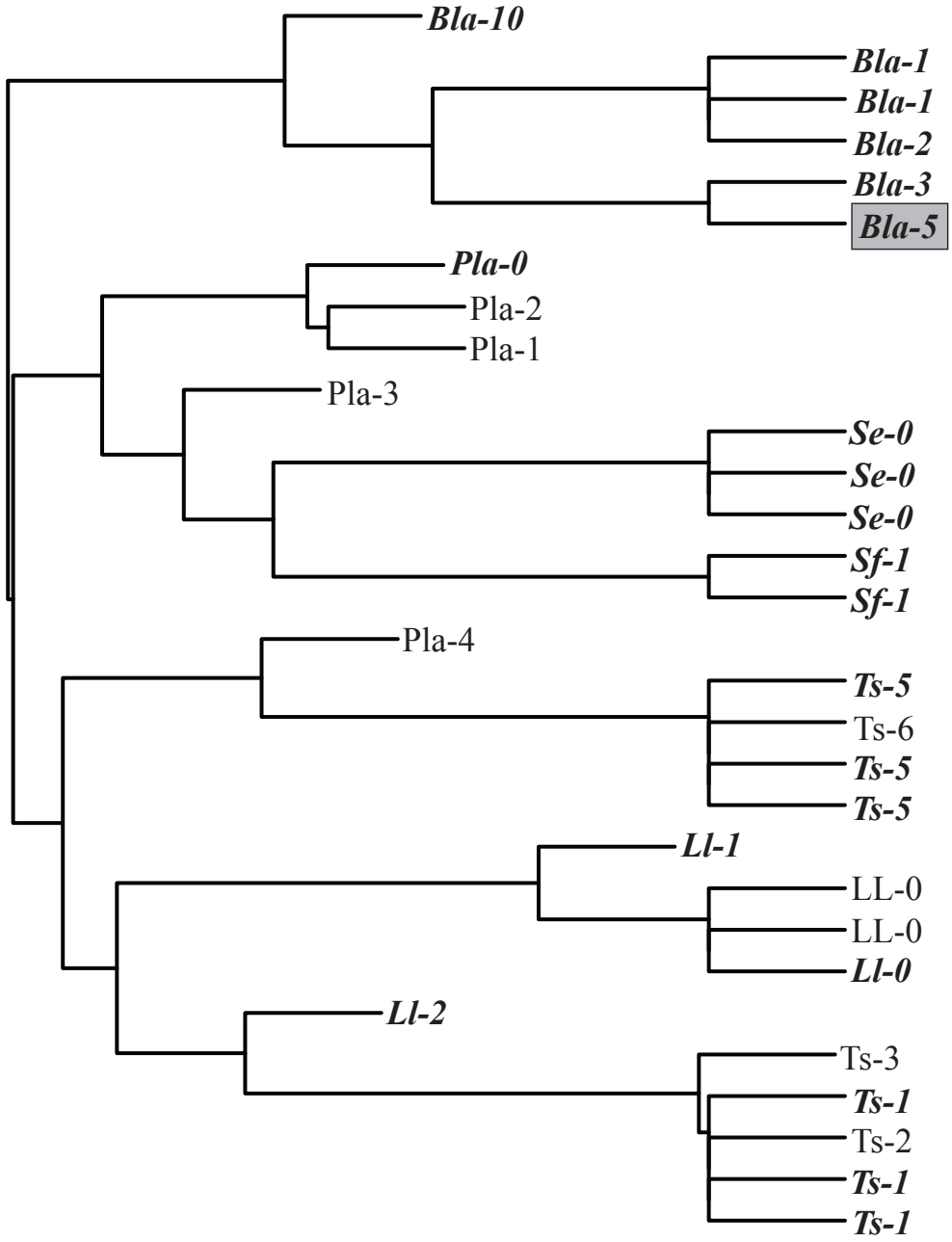

Figure 1.7. Overlay of ploidy and genetic distance for Bal-5 and genetically similar accessions. This subclade is derived from a larger tree consisting of 5964 accessions based on 139 SNPs. The tree was generated in R using Hierarchical Clustering. Ploidy was determined by flow cytometric analysis of isolated nuclei for all accessions in bold italic. Untested accessions are in normal text, confirmed diploids are in bold italic and tetraploids are boxed.

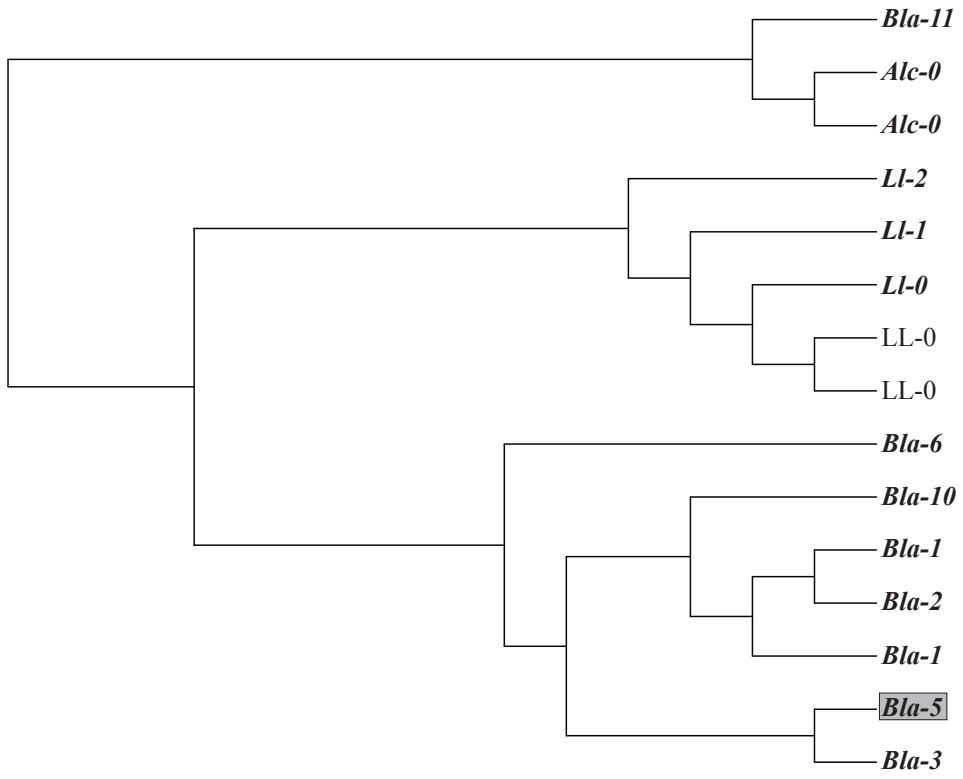

Figure 1.8. Overlay of ploidy and genetic distance for Bal-5 and genetically similar accessions. This subclade is derived from a larger tree consisting of 5964 accessions based on 139 SNPs. The distance matrix was calculated using the jukes-cantor method and the tree was calculated using neighbor joining. Ploidy was determined by flow cytometric analysis of isolated nuclei for all accessions in bold italic. Untested accessions are in normal text, confirmed diploids are in bold italic and tetraploids are boxed.

Table 1.4. The frequencies of polymorphisms between accessions in the Swedish clade among 139 SNPs.

	M3385S	**M7943S**	**Ste-3**	**T510**
M3385S	0	0	0.504	0.381
M7943S		0	0.504	0.381
Ste-3			0	0.512
T510				0

Table 1.5. The frequencies of polymorphisms between accessions in the Blanes clade based on 139 SNPs.

	Bla-5	**Bla-3**	**Bla-2**	**Bla-1 (CS6616)**	**Bla-1 (N971)**
Bla-5	0	0.014	0.187	0.187	0.180
Bla-3		0	0.180	0.180	0.180
Bla-2			0	0	0.022
Bla-1 (CS6616)				0	0.022
Bla-1 (N971)					0

an established tetraploid population we would expect that Bla-3, Bla-1, and Bla-2, and perhaps even Bla-10 and Bla-6 would be tetraploids also. We determined through flow cytometry that the ploidy level of Bla-3 and many of the other accessions in the clade were diploid (Figure 1.7). This was different from Wa-1 and M3385S because this tetraploid had a clear congenic diploid sister accession that may represent the diploid parental population of Bla-5. Either Bla-5 was a single tetraploid from the diploid population or the tetraploid population has not had enough time to diverge from the diploid population.

To determine if Bla-5 represents a tetraploid population the ploidy level of multiple *A. thaliana* accessions collected from costal Spain were measured (Table 1.1). Assuming that random collections over the same area would find other tetraploids from an established population, nuclear DNA contents of *A. thaliana* accessions from Tossa de Mar, Llagoster, La Montagoda, and other locations were analyzed. None of the plants from these regions were tetraploids, consistent with Bla-5 being derived from a rare off-type tetraploid and not a persistent adapted tetraploid population.

1.3.5 Ciste-2 a newly detected tetraploid with detailed provenance from Cisterna di Latina, Italy

Flow cytometry was used to measure the ploidy level of a small group of the first 80 accessions with released resequenced genomes from the 1001 genomes project (www.1001genomes.org). Of these only one, Ciste-2, was tetraploid, which confirms previous results (Chao et al. 2013). Unlike the other tetraploids in this study the precise collection site of Ciste-2, in Cisterna di Latina, Latina, Italy (latitude 41.615583º N, longitude 12.868655º E), was provided by the Arabidopsis Biological Resource Center

(ABRC) at The Ohio State University (http://arabidopsis.org/abrc/index.jsp). Ciste-2 was collected at a site that is also inhabited by a diploid accession, Ciste-1, according to ABRC.

Ciste-2 was not genotyped as part of the species wide diversity study. The sequence data was available, however, and we extracted the 139 SNP positions from the alignment of Ciste-2 sequence to the Col-0 reference (Cao et al. 2011). Figures 1.9 and 1.10 showed that Ciste-2 was most closely related to ecotypes in Italy, Croatia, and Austria. Using flow cytometry, we determined that all genetically similar accessions were diploid (Table 1.1)**.**

To test if Ciste-2 was genetically similar to any of these diploids, we compared Ciste-2 SNPs to the other accessions in the cluster generated by neighbor-joining. Ciste-2 did not share many polymorphisms with these accessions and there was no evidence that Ciste-2 was derived from one of these populations (Table 1.6). Since Ciste-2 was not genetically similar to Ciste-1 or any other diploid, it could be a tetraploid population cohabiting the same area as diploid populations. However, flow cytometric analysis of individuals collected from that region by Doug Schemske in 2011 indicated that all were diploid (personal correspondence). The failure to find another tetraploid indicates that a thriving tetraploid population in the area does not exist. Thus, there is no evidence to support the hypothesis that Ciste-2 represents a persistent tetraploid population.

1.4 Discussion

Polyploidy or ancient polyploidy is evident in every angiosperm sequenced to date, suggesting that polyploidy has had a major influence on the evolution and

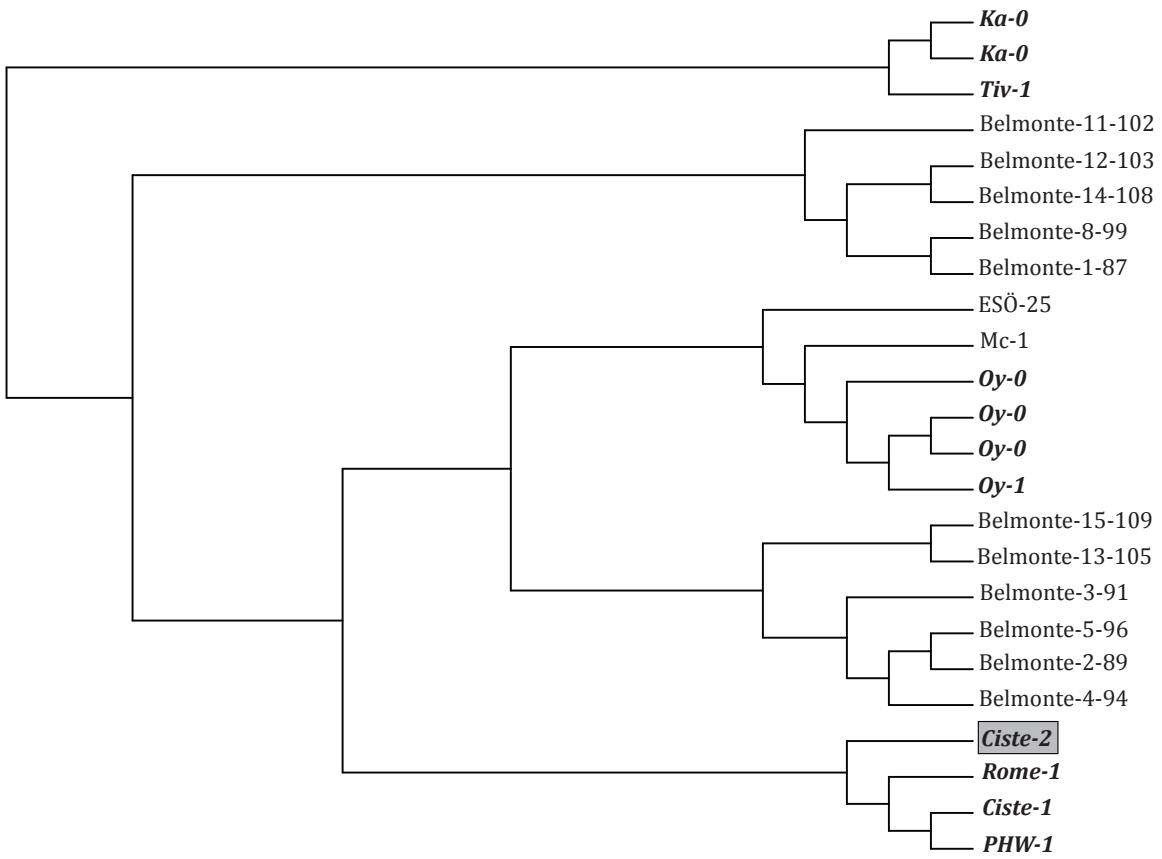

Figure 1.9. Overlay of ploidy and genetic distance for Ciste-2 and genetically similar accessions. This subclade is derived from a larger tree consisting of 5964 accessions based on 139 SNPs. The tree was generated in R using Hierarchical Clustering. Ploidy was determined by flow cytometric analysis of isolated nuclei for all accessions in bold italic. Untested accessions are in normal text, confirmed diploids are in bold italic and tetraploids are boxed.

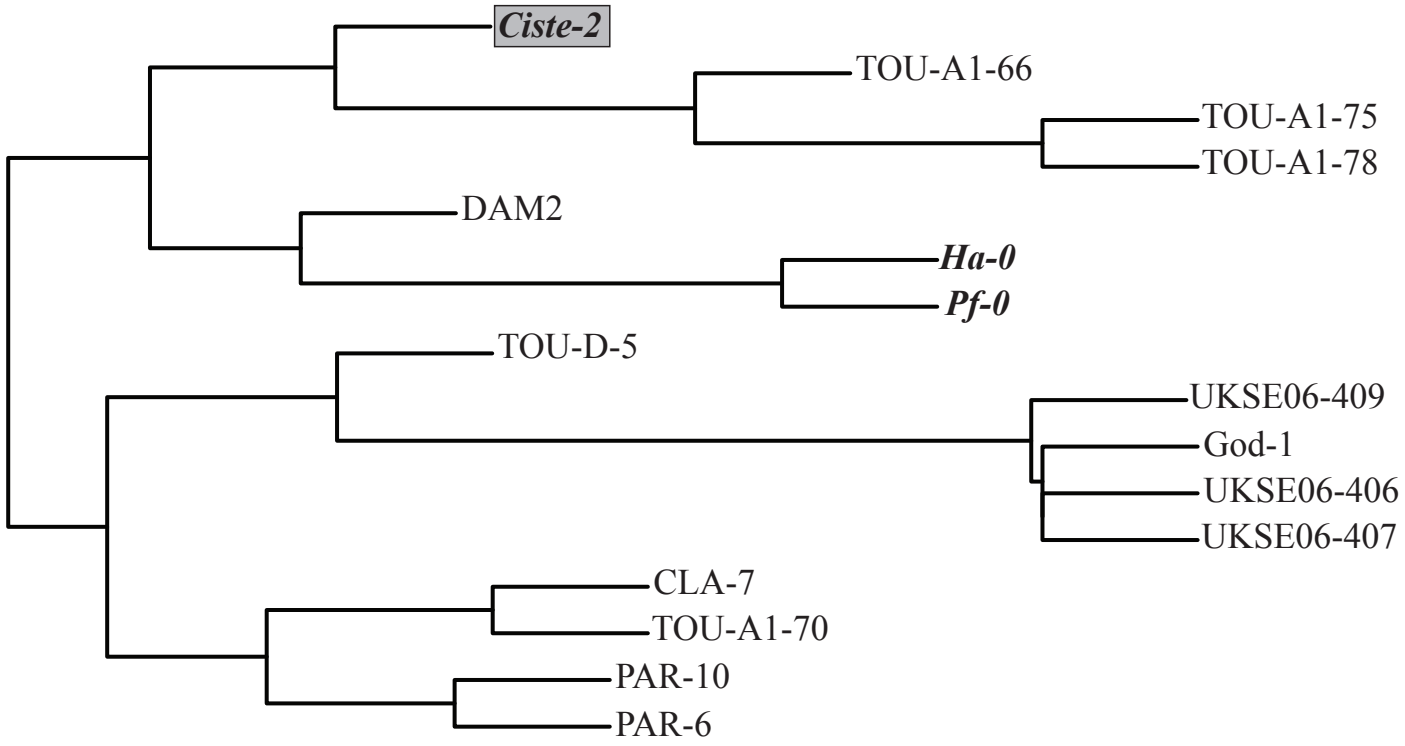

Figure 1.10. Overlay of ploidy and genetic distance for Ciste-2 and genetically similar accessions. This subclade is derived from a larger tree consisting of 5964 accessions based on 139 SNPs. The distance matrix was calculated using the jukes-cantor method and the tree was calculated using neighbor joining. Ploidy was determined by flow cytometric analysis of isolated nuclei for all accessions in bold italic. Untested accessions are in normal text, confirmed diploids are in bold italic and tetraploids are boxed.

Table 1.6. The frequencies of polymorphisms between accessions in the Cisterna di Latina clade based on 139 SNPs.

	Ciste-2	**TOU-A1-66**	**TOU-A1-75**	**TOU-A1-78**
Ciste-2	0	0.396	0.345	0.345
TOU-A1-66		0	0.194	0.179
TOU-A1-75			0	0.014
TOU-A1-78				0

diversification of the angiosperms. Since polyploidy continues to appear in contemporary plant populations it is sometimes presumed that their existence demonstrates that there must be some scenarios that provide an advantage for polyploids over their diploid progenitors. The selection providing this advantage should also drive divergence of the new population from the progenitor diploid population resulting in strong population differentiation after a period of successful tetraploid reproduction.

One such potential advantage is the loss of constraint, due to higher allele copy number, that should also permit the accumulation of new variants without selection and even the accumulation of weak deleterious alleles, especially in a plant with a predominant selfing reproductive habit. However, autopolyploidy also has disadvantages such as disrupted meiosis leading to aneuploidy. We would expect that selection for alleles that compensate for the negative consequences of polyploidy and better adapt an established tetraploid to the effects of polyploidy would also contribute to divergence from a diploid progenitor (Hollister et al. 2012).

A. thaliana does produce 2n gametes and produces a low rate of tetraploids arising from diploid populations. Abiotic stresses increase the frequency of unreduced gametes (Ramsey and Schemske 1998, De Storme et al. 2012). For example, De Storme et al. (2012) cold shocked the plants at 4ºC to 5º C and found that many pollen spores were triploid or tetraploid. They hypothesized that cold stress is a natural mechanism that may lead to polyploidization in natural populations. The average minimum temperature in Sweden and Poland in May is less than 10º C (www.weather-and-climate.com), and potentially flowering plants do experience cold stress and high frequency of unreduced gametes occur in Swedish and Polish diploid populations. Nonetheless, an established *A.*

thaliana population would only occur if the tetraploids had an advantage over the diploid or the two cytotypes were reproductively isolated from each other (Levin 1975, Ramsey and Schemske 1998, Husband 2000). It was unclear if *A. thaliana* actually had natural tetraploid populations; therefore, it was unclear if tetraploidy was a characteristic that contributed to the diversification of the species, or if tetraploidy was a transient character state that did occur, but potentially had neither advantage for the species, nor contribute any role to its diversification and local adaptation.

To test whether *A. thaliana* is a transient or persistent character state we looked at the distribution of tetraploids and the genetic diversity of tetraploids using four known tetraploids. First, we determined that the tetraploids were independently derived since they share no common collection site and they are not genetically similar (Figures 1.1 & 1.2). Next, we determined whether each individual tetraploid belonged to a tetraploid population. We hypothesized that either these tetraploids were sub-tetraploid populations and other closely related accessions should also be tetraploids, or they were singular tetraploid events that have been captured and propagated in the seed stock centers (Platt et al. 2010, Bomblies et al. 2010). To determine if the tetraploids were collected from tetraploid populations we measured the ploidy level of the accessions most closely related to each tetraploid and calculated the genetic similarities between the ecotypes.

Wa-1 and Me-0 were the same ecotype, and it was the only tetraploid within the respective clade (Figures 1.3 & 1.4). The collection site of this ecotype was unknown because its collection and storing history were not documented. Most likely, Wa-1 and Me-0 represent a single collection by Albert Kranz followed by strain duplications and labeling errors during culture (Platt et al. 2010, Anastasio et al. 2011). Also, the two

proposed collection sites from Albert Kranz is over 850 km, which is a distance too great to maintain the high genetic similarities observed between these two accessions (Platt et al. 2010). Clearly, the region where Wa-1/Me-0 were collected is not well resampled, as the other diploids within its clade come from different countries, and without resampling we do not know for sure what the ploidy level of the Wa-1/Me-0 natural population was. However, based on the current data, there is no evidence that Wa-1/Me-0 came from a persistent tetraploid population.

M3385S and M7943S represented another unique tetraploid since these two accessions were genetically identical (Figures 1.5 & 1.6). M3385S and M7943S are Swedish accessions, but their exact collection sites are unknown. M3385S and M7943S were significantly diverged from the most closely related diploids, M7884S and H55/M7323S, and share the same collector and provenance. However, the clade appears to be lacking in sampling also, as all individuals in the clade are very genetically diverged from each other. This also would explain the differences between the trees generated by hierarchical clustering and neighbor-joining. The sampling was sparse in the region where the tetraploid ecotype came from and therefore, the tetraploid did not cluster well with the other accessions. It is interesting to note that the hierarchical clustering tree clustered the tetraploid with some accessions from Italy and the neighbor-joining tree clustered the Swedish tetraploid with Danish and German diploids. Some of the German diploids appeared in the hierarchical cluster tree of Ciste-2; whereas, the neighbor-joining tree had a Swedish/Norwegian clade closely related to the Ciste-2.

The tetraploid, Bla-5, was genetically identical to the diploid, Bla-3 (Table 1.5). Either Bla-5 is an example of a rare tetraploid that was collected in an otherwise diploid

population, or it was collected from a tetraploid population that has not had significant time to diverge from the diploid population. We can rule out the latter hypothesis, because unlike the Wa-1 and the M3385S, Blanes, Spain has been thoroughly sampled. In addition, similar ecological contexts along the Costa Brava in Spain have been thoroughly sampled by multiple researchers and in all of those tested to date all but Bla-5 are diploid. The frequent sampling along the Costa Brava may owe in part to the interest in Iberian subpopulations and salt adaptation in Arabidopsis but surely is also due to the higher frequency with which the Spanish Mediterranean coast was used as a vacation destination for Arabidopsis researchers as compared to Poland, rural Germany, and rural inland Sweden. Work to expand collections in Fennoscandia may help alleviate some of this under sampling.

Ciste-2 is also another tetraploid that can be traced back to the collection site, and find a diploid population. Like Bla-5, there was no more evidence of a tetraploid population in that region. Latina, Italy and other parts of Italy have been well sampled for *A. thaliana* populations, and Ciste-2 is the only tetraploid to be found.

Based on the lack of support for established tetraploid populations we propose that tetraploidy is a transient state in *A. thaliana.* Though tetraploidy can and does arise in populations of *A. thaliana* we find no evidence that tetraploids establish lasting populations and stably reproduce as tetraploids. Therefore, we propose that it is not an appropriate model for studying the evolution and population dynamics of polyploidy since tetraploidy is not a persistent character state. Another model system should be selected that is more appropriate for studying the effects of polyploidy on population dynamics and for studying the mechanisms that govern the establishment of polyploidy in

the wild. The outcrossing sister species, *A. arenosa*, exists as both diploid and tetraploid cytotypes with some geographical differentiation (Hollister et al. 2012). Since the biological question is available in this species it may be better suited for these studies.

1.5 References

Anastasio, AE, A Platt, M Horton, E Grotewold, R Scholl, JO Borevitz, M Nordborg, J Bergelson (2011) Source verification of mis-identified Arabidopsis thaliana accessions. Plant J 67:554–66

Bergelson, J, E Stahl, S Dudek, M Kreitman (1998) Genetic Variation Within and Among Populations of Arabidopsis thaliana. Genetics 148:1311–1323

Birchler, JA, H Yao, S Chudalayandi, D Vaiman, RA Veitia (2010) Heterosis. Plant Cell 22:2105–12

Bomblies, K, J Lempe, P Epple, N Warthmann, C Lanz, JL Dangl, D Weigel (2007) Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. PLoS Biol 5:e236

Bomblies, K, L Yant, R a Laitinen, S-T Kim, JD Hollister, N Warthmann, J Fitz, D Weigel (2010) Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. PLoS Genet 6:e1000890

Cao, J, K Schneeberger, S Ossowski, T Günther, S Bender, J Fitz, D Koenig, C Lanz, O Stegle, C Lippert, X Wang, F Ott, J Müller, C Alonso-Blanco, K Borgwardt, KJ Schmid, D Weigel (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43:956–63

Chao, D-Y, B Dilkes, H Luo, A Douglas, E Yakubova, B Lahner, DE Salt (2013) Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis. Science 341:658–9

Comai, L (2005) The advantages and disadvantages of being polyploid. Nat Rev Genet 6:836–46

Comai, L, a P Tyagi, K Winter, R Holmes-Davis, SH Reynolds, Y Stevens, B Byers (2000) Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. Plant Cell 12:1551–68

Cui, L, PK Wall, JH Leebens-Mack, BG Lindsay, DE Soltis, JJ Doyle, PS Soltis, JE Carlson, K Arumuganathan, A Barakat, V a Albert, H Ma, CW dePamphilis (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res 16:738–49

Dilkes, BP, M Spielman, R Weizbauer, B Watson, D Burkart-Waco, RJ Scott, L Comai (2008) The maternally expressed WRKY transcription factor TTG2 controls lethality in interploidy crosses of Arabidopsis. PLoS Biol 6:2707–20

Doyle, GG (1986) Aneuploidy and inbreeding depression in random mating and self-fertilizing autotetraploid populations. Theor Appl Genet 72:799–806

Felsenstein, J (2004) PHYLIP (Phylogeny Inference Package). Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Freeling, M, BC Thomas (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res 16:805–14

Henry, IM, BP Dilkes, L Comai (2006) Molecular karyotyping and aneuploidy detection in Arabidopsis thaliana using quantitative fluorescent polymerase chain reaction. Plant J 48:307–19

Henry, IM, BP Dilkes, a P Tyagi, H-Y Lin, L Comai (2009) Dosage and parent-of-origin effects shaping aneuploid swarms in A. thaliana. Heredity (Edinb) 103:458–68

Henry, IM, BP Dilkes, K Young, B Watson, H Wu, L Comai (2005) Aneuploidy and genetic variation in the Arabidopsis thaliana triploid response. Genetics 170:1979–88

Hollister, JD, BJ Arnold, E Svedin, KS Xue, BP Dilkes, K Bomblies (2012) Genetic adaptation associated with genome-doubling in autotetraploid Arabidopsis arenosa. PLoS Genet 8:e1003093

Horton, MW, AM Hancock, YS Huang, C Toomajian, S Atwell, A Auton, NW Muliyati, A Platt, FG Sperone, BJ Vilhjálmsson, M Nordborg, JO Borevitz, J Bergelson (2012) Genome-wid patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel 44:212–216

Husband, BC (2000) Constraints on polyploid evolution : a test of the minority cytotype exclusion principle. Proc R Soc London, B, Biol Sci 267:217–223

Levin, DA (1975) Minority Cytotype Exclusion in Local Plant Populations. Taxon 24:35–43

Li, W-L, GP Berlyn, PMS Ashton (1996) POLYPLOIDS AND THEIR STRUCTURAL AND PHYSIOLOGICAL CHARACTERISTICS RELATIVE TO WATER DEFICIT IN BETULA PAPYRIFERA (BETULACEAE). Am J Bot 83:15–20

Lynch, M, AG Force (2000) The Origin of Interspecific Genomic Incompatibility via Gene Duplication. Am Nat 156:590–605

Müntzing, A (1951) Cyto-genetic properties and practical value of tetraploid rye. Hereditas 37:17–84

Ohno, S (1970) Evolution of Gene Duplication. New York: Springer-Verlag

Otto, SP, J Whitton (2000) Polyploid Incidence and Evolution. Annu Rev Genet 34:401–437

Petit, C, F Bretagnolle, F Felber (1999) Evolutionary consequences of diploid–polyploid hybrid zones in wild species. Trends Ecol Evol 14:306–311

Pignatta, D, BP Dilkes, S-Y Yoo, IM Henry, A Madlung, RW Doerge, Z Jeffrey Chen, L Comai (2010) Differential sensitivity of the Arabidopsis thaliana transcriptome and enhancers to the effects of genome doubling. New Phytol 186:194–206

Platt, A, M Horton, YS Huang, Y Li, AE Anastasio, NW Mulyati, J Agren, O Bossdorf, D Byers, K Donohue, M Dunning, EB Holub, A Hudson, V Le Corre, O Loudet, F Roux, N Warthmann, D Weigel, L Rivero, R Scholl, M Nordborg, J Bergelson, JO Borevitz (2010) The scale of population structure in Arabidopsis thaliana. PLoS Genet 6:e1000843

R Core Team (2013) R: A language and environment for statistical computing. Vienna, Austria

Ramsey, J, DW Schemske (1998) Pathways, mechanisms and rates of polyploid formation in flowering plants. Annu Rev Ecol Syst 29:467–501

Randolph, LF (1935) Cytogenetics of tetraploid maize. J Agric Res 50:591–605

Rodriguez, DJ (1996) A Model for the Establishment of Polyploidy in Plants. Am Nat 147:33–46

Saleh, B, T Allario, D Dambier, P Ollitrault, R Morillon (2008) Tetraploid citrus rootstocks are more tolerant to salt stress than diploid. C R Biol 331:703–10

Schmuths, H, A Meister, R Horres, K Bachmann (2004) Genome size variation among accessions of Arabidopsis thaliana. Ann Bot 93:317–21

Soltis, DE, RJA Buggs, JJ Doyle, PS Soltis (2010) What we still don't know about polyploidy. Taxon 59:1387–1403

Soltis, PS, DE Soltis (2000) The role of genetic and genomic attributes in the success of polyploids. Proc Natl Acad Sci U S A 97:7051–7

De Storme, N, GP Copenhaver, D Geelen (2012) Production of diploid male gametes in Arabidopsis by cold-induced destabilization of postmeiotic radial microtubule arrays. Plant Physiol 160:1808–26

Weiss, H, J Maluszynska (2000) Chromosomal rearrangement in autotetraploid plants of Arabidopsis thaliana. Hereditas 133:255–261

Wood, TE, N Takebayashi, MS Barker, I Mayrose, PB Greenspoon, LH Rieseberg (2009) The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci U S A 106:13875–9

Yu, Z, K Haage, VE Streit, A Gierl, RAT Ruiz (2009) A large number of tetraploid Arabidopsis thaliana lines, generated by a rapid strategy, reveal high stability of neo-tetraploids during consecutive generations. Theor Appl Genet 118:1107–1119

CHAPTER 2. PIPELINE LINKING GENOTYPE TO PHENOTYPE USING GENOME-WIDE ASSOCIATION

2.1 Introduction

Genome-wide association (GWA) has been a common tool used to undercover causative genes of human traits and more recently being used to discover causative genes of plant phenotypes (Zeggini et al. 2007, The Wellcome Trust Case Control Consortium 2007, Todd et al. 2007, Chan et al. 2010, Atwell et al. 2010, Tian et al. 2011). GWA studies were first developed as an alternative to quantitative trait loci (QTL) analyses because QTL studies require inbred lines to find causative polymorphisms within a population. QTL analyses were not possible in humans, but GWA studies calculate the natural variance of a population to map phenotype to genotype, which was more ideal for the human population. The natural variation was mapped by using markers, such as single-nucleotide polymorphisms (SNPs). Once GWA showed success in the human population, the method started being used for other organisms such as *Arabidopsis thaliana* (Chan et al. 2010, Atwell et al. 2010, Li et al. 2010).

Different statistical models have been developed to calculate the association between SNPs and phenotype while accounting for population structure (Devlin and Roeder 1999, Pritchard et al. 2000, Price et al. 2006, Patterson et al. 2006, The Wellcome Trust Case Control Consortium 2007, Kang et al. 2008, 2010, Sabatti et al. 2009,

Cho et al. 2009). The general linear mixed model performs better than other methods because it incorporates the pairwise genetic relatedness of the individuals in every comparison between SNP and phenotype (Yu et al. 2006, Zhao et al. 2007, Malosetti et al. 2007, Kang et al. 2008). It also incorporates a variance component that models the phenotypic correlation between individuals (Kang et al. 2010). The statistical mixed model is

$$y = X\beta + Zu + e$$

where y is the association between genotype and phenotype. X is a matrix of the fixed effects, which are the genotype markers. β is a matrix of the coefficients of the fixed effects. Z is an incidence matrix mapping each phenotype with an individual. u is the random effect, and e is a matrix of random error (Yu et al. 2006, Kang et al. 2008). The variance of u is calculated as $\mathrm{Var}(u) = \sigma^2_g K$, where K represents the kinship matrix. The variance-covariance matrix of the phenotype is calculated as $V = \sigma^2_g ZKZ' + \sigma^2_e I$ (Kang et al. 2008). The linear mixed model has fewer spurious positives and more power than other methods (Yu et al. 2006, Zhao et al. 2007, Malosetti et al. 2007, Kang et al. 2008).

The calculations required for genome associations using the mixed model can be computationally intensive and require several hours or days for completing large data sets. One method called efficient-mixed model association (EMMA) was developed to decrease the computational time needed to complete the GWA (Kang et al. 2008). EMMA uses phylogenetic control to calculate a simple genetic matrix to be used as the kinship matrix. The phylogenetic control assumes that the phylogenetic tree of the population is a good approximation of the relatedness of the population (Kang et al. 2008). However, EMMA is still too computationally intensive for very large data sets

since it recalculates the variance parameters for each SNP. Kang et al. (2010) developed a modified version of EMMA called EMMA eXpedited (EMMAX). EMMAX uses the phylogenetic control model to calculate the kinship file, however, instead of recalculating the variance component matrix for each association it estimates the variance parameters once, globally applying the parameters to each association, assuming that many genes contribute to a specific trait. This allowed for a faster calculation time and therefore, more SNPs and individuals could be included in the analysis (Kang et al. 2010).

A second method, called multi-locus mixed-model (MLMM), was created using EMMA as a foundation (Segura et al. 2012). MLMM recalculates the variance parameters at each step like EMMA, however MLMM differs from EMMA by including multiple loci into the model. MLMM uses simple stepwise mixed-model to determine the best model for the phenotype. Segura et al. (2012) hypothesized that multiple loci had associations with a trait only because of linkage disequilibrium (population structure), and by including SNPs into the model the spurious positives due to population structure would be eliminated (Segura et al. 2012). By including multiple loci into the model, the false discovery rate (FDR; discussed below) is lower and power is higher than a single-locus model test (Segura et al. 2012).

MLMM uses two different model-selecting criteria to select the correct association model: extended Bayesian Information Criterion (EBIC) and Bonferroni. Segura et al. (2012) showed that EBIC was a more stringent method than Bonferroni for selecting a model, but both performed well and the FDR was low. MLMM significantly eliminates the number of significant SNPs, and the output of MLMM includes only the

significant SNPs according to the EBIC and Bonferroni significant cutoffs (Segura et al. 2012).

The standard significant cutoff for a single hypothesis test is $\alpha \leq 0.05$. During GWA, thousands or millions of hypotheses are tested, depending on the number of SNPs used in the study; therefore, to correct for the number of hypotheses tested, a new significant cutoff is calculated. A simple Bonferroni correction is

$$\alpha \leq \frac{0.05}{\#\ \text{hypotheses}}$$

For 211K SNPs, the Bonferroni cutoff would be $\alpha \leq 2.37\text{x}10^{-7}$. For 4.9 million SNPs the new Bonferroni cutoff would be $\alpha \leq 1.02\text{x}10^{-8}$. However, the Bonferroni method is one of the most stringent correction tests. Another method for correcting for the number of hypotheses tested is to calculate the FDR.

To eliminate spurious positive associations the FDR is used to calculate a new significant threshold. It is less stringent than Bonferroni (Benjamini and Hochberg 1995, 2000, Storey 2002, Verhoeven et al. 2005, Benjamini et al. 2006, Pike 2011). The Benjamini-Hochberg method for calculating the FDR required ranking of the *p*-values and determining a new significant cutoff based on the rank (Benjamini and Hochberg 1995, Pike 2011). The equation for Benjamini-Hochberg FDR is

$$\alpha \leq \frac{0.05 * \text{rank}}{\#\ \text{hypotheses}}$$

where the rank is 1,2,…,*n* and *n* is the total number of hypotheses being tested. The very first test is a Bonferroni test, but the second test multiples 0.5 by 2 and thus increases α, the new significant threshold (Benjamini and Hochberg 1995, Pike 2011). Ranking the *p*-

values allows for a less stringent method of calculating the false discovery rate than Bonferroni.

Another method to deflate the number of spurious positives is to eliminate SNPs with minor allele frequencies (MAF) ≤ 5 or 10% from the genotype data set (Florez et al. 2007, Cupples et al. 2007, Wray et al. 2011, Freudenberg et al. 2011). It is hypothesized that MAF ≤ 10% will bias the hypothesis test because the chance of individuals sharing an allele having the same phenotype is greater than if MAF ≥ 25%. However, Tabangin et al. (2006) showed that SNPS with MAF ≤ 10% did not result in more spurious positives than was expected. They concluded that it would be erroneous to remove those SNPs from the data set since the causal SNP might have a MAF ≤10% (Tabangin et al. 2009, Abdollahi-Arpanahi et al. 2014). My pipeline includes SNP data sets containing all SNPs, and also data set that have the SNPs with MAF ≤ 5% removed.

2.2 Genetic datasets

We created two different genotype files. The first file contains 211,781 SNPs compiled from sequence data of 80 accessions and SNPs called for 360 accessions (Platt et al. 2010, Atwell et al. 2010, Cao et al. 2011, Horton et al. 2012) for a total of 428 accessions. This data set is referred to as the 211K SNPs data set. The second file contains 4.9 million SNPs for 430 accessions generated using the sequence data available from the 1001genomes project and imputation of missing data. This data set is referred to as the 4.9M SNPs data set. The master SNP file created contained 466 accessions, 244 accessions had the complete 4.9 million SNPs from sequence data with various SNPs missing in each accession, and 222 accessions only had 211K SNPs from the array data.

The missing data for the 222 accessions and the random missing data for the 244 accessions were imputed using BEAGLE.

These two genotype files were formatted as tped files. Each row was a SNP and each accession was represented in two columns (Figure 2.1). EMMAX used tped file format. To run MLMM, the files were transposed and each accession only represented once; therefore, each accession was a single row and each SNP was a column (Figure 2.2).

An additional SNP file was created for the 211K SNPs and the 4.9M SNPs data sets by eliminating any SNP with a low MAF (MAF ≤ 5%). The 211K SNPs data set had 198,409 SNPs after eliminating the low allele frequency SNPs. The 4.9M SNPs data set had 1.6 million (1.6M) SNPs after eliminating the low allele frequency SNPs.

2.3 Pipeline

We created a pipeline that uses one of the two genotype files and a phenotype file containing as many phenotypes desired. The output of the pipeline includes manhattan plots (used to visualize the output) for every phenotype, significant SNPs for every phenotype, and lists of genes linked to each significant SNP. The pipeline was created so that running a genome-wide association for multiple traits was time-efficient by only requiring one initial start-up for a set of phenotypes.

The scripts were written in Perl and executed using Unix in a command window. EMMAX requires five scripts: Running_GWAS.pl, ManhanFiles_Plots.pl, FindingSignificantSNPs.pl, CalculateFDR.pl, and Determine_Genes_LinkedToSigSNPs.pl. These scripts can be found in Appendices A-E, respectively. MLMM requires only Perl two scripts because an R script was provided to

1 1657 0 657 1 1 1 1 1 1 1 1 2 2 2 2 1 1 1 1 1 1 1...
1 13102 0 3102 1 1 1 1 1 1 1 1 2 2 2 2 1 1 1 1 1 1 1...
1 14648 0 4648 1 1 1 1 1 1 1 1 2 2 2 2 1 1 1 1 1 1 1...
1 14880 0 4880 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1...
1 15975 0 5975 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1...
1 16063 0 6063 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2...
1 16449 0 6449 1 1 1 1 1 1 2 2 1 1 1 1 2 2 2 2 1 1 2...
1 16514 0 6514 1 1 1 1 1 1 1 1 2 2 2 2 1 1 1 1 1 1 1...
1 16603 0 6603 1 1 1 1 1 1 1 1 2 2 2 2 1 1 1 1 1 1 1...
1 16768 0 6768 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1...
...

Figure 2.1. Example of the tped file used for EMMAX. Each SNP was represented in a row, and each accession was represented by two columns. The columns were as follows: Chromosome, SNP ID, Dummy family variable, SNP position, and the remaining columns are the SNP dataset. SNPs were represented by numbers: 1=reference SNP, 2=non-reference SNP, and 0=missing data.

Line,1- 10969,1- 12584,1- 12659,1- 13045,1- 14385,1- 20892,1- 21043,1- 21128,1- 23838,1- 25315,...
Col-0,0,0,0,0,0,0,0,0,0,0
RRS-10,0,0,0,0,0,0,0,0,0,0
Aa-0,1,1,0,0,1,1,1,1,0,1
Alst-1,0,0,0,0,0,0,1,1,0,0
Amel-1,0,1,0,1,0,0,1,0,0,0
An-2,0,1,0,1,0,0,1,0,0,0
Ang-0,0,1,0,0,0,0,1,1,0,0
Ann-1,0,1,0,0,0,0,1,1,0,0
Arby-1,0,0,0,0,0,0,0,0,0,0
Ba-1,0,1,0,0,0,0,1,1,0,0
...

Figure 2.2 Example of SNP file required for MLMM. Each accession is represented by one row and each SNP is represented by one column. The first column is the name of the accessions and the remaining columns are the SNPs named by Chr- SNP position.

run MLMM and the significant SNPs and manhattan plots are produced using the R script. However, we modified the R script to produce more output and the modified R script can be found in Appendix F. The two Perl scripts for the MLMM output are CreatingSignificantSNPFiles.pl and DetermineGenes_LinkedToSigSNPs.pl. The script CreatingSignificantSNPFiles.pl can be found in Appendix G.

2.3.1 Running EMMAX

The initial script used to run EMMAX uses the phenotype file, genotype file, and the software required to run EMMAX. The kinship matrix was calculated as instructed on the EMMAX website (http://genetics.cs.ucla.edu/emmax/).

The Perl script used is called Running_GWAS.pl. The script parsed the phenotype file so that only one phenotype is done at a time. Once the phenotype file is parsed into the different phenotypes, a new file was created containing only one phenotype and was named phenotype.phenos. The file was saved into a new folder with the name of the phenotype. For example if the first phenotype in the phenotype file was named PlantHeight the new phenotype file would be named PlantHeight.phenos and saved in a folder called PlantHeight. The command to run EMMAX was executed for the specific phenotype and saved in the respective phenotype folder. This was reiterated until the completion of all the phenotypes in the phenotype file. The EMMAX ouput includes 3 folders: phenotype.log, which contains the EMMAX calculations; phenotype.reml, which contains the residual maximum likelihood (REML) estimates of the fit of the model, and phenotype.ps, which contains the *p*-values and β-values of the association of the SNPs and the phenotype. The .ps file is used for the remainder of the pipeline.

2.3.1.1 Creating manhattan plots

Visualizing the output of EMMAX is helpful in understanding the output, and getting a clear picture of the loci showing high significance. Manhattan plots are the common form of visualizing the GWA output. Manhattan plots are created using R using an R script that we modified to help visualize specific SNPs in the manhattan plots (Turner 2011, R Core Team 2013). The original script allowed for highlighting specific SNPs with one color, but the script was modified to allow for highlighting different groups of specific SNPs in different colors. This was used to highlight different genes of interest. The modified script is available in Appendix H.

The Perl script used to ease the process of creating manhattan plots for every phenotype is called ManhanFiles_Plots.pl. It opens each .ps file and modifies the format of the file and names the columns. The file format required to run the manhattan command in R was the first column being the SNP id, the second column was the chromosome, the third column was the base pair position, the fourth column was the *p*-value, and the fifth column was the β-value. The fifth column was optional because it is not used to create the manhattan plots. We added it so that no information was lost in the file modifications. The columns were named as the following: SNP CHR BP P B. The file was named phenotype.manhan.csv and saved in a newly created folder called "manhattanfiles." In addition an R script, named "readmanhanfilesintoR.r" was printed that contains the script required for reading in the .manhan.csv files, running the manhattan command in R, and saving the images. The R script needed to be copied and pasted into the R console.

The R script requires the home directory of the .manhan.csv files. This can be changed either in the Perl script, ManhanFiles_Plots.pl or can be changed by finding and replacing in the "readmanhanfilesintoR.r" script, if the home directory of the files changes during the analysis. Before running the "readmanhanfilesintoR.r" script in R, the manhattan R script needed to be loaded.

The images saved are either .jpeg or .pdf depending on the preference of the user. To change the image format, the "ManhanFiles_Plots.pl" script needs to be changed. The name of the image will either be .jpg or .pdf. The actual command to create the image will either be jpeg() or pdf(). This can be found on line 127 of the script.

Line 127: print RSCRIPT "dev.print(device=postscript, \"$folders[$f].**jpg**\", onefile=FALSE,horizontal=FALSE);\n**jpeg**(\"$folders[$f].**jpg**\");\n manhattan($folders[$f],pch=20,main=\"$folders[$f]\");\ndev.off()\n\n";

or

Line 127: print RSCRIPT "dev.print(device=postscript, \"$folders[$f].**pdf**\", onefile=FALSE,horizontal=FALSE);\n**pdf**(\"$folders[$f].**pdf**\");\n manhattan($folders[$f],pch=20,main=\"$folders[$f]\");\ndev.off()\n\n";

To annotate different genes the original command is

manhattan(phenotype, annotate=SNP List).

The modifications added the option to annotate multiple SNP clusters with different colors. Thus, the new command is

manhattan(phenotype, annotate**1**=SNP List1, annotate**2**=SNP List2…)

Table 2.1 contains the annotation codes and their correlating colors. 13 different SNP clusters, e.g. genes, can be annotated with different colors. Additional colors may be added by modifying the manhattan script.

Table 2.1 The coding for highlighting SNPs in a manhattan plot and the colors that correspond to each annotation code.

Code	Color Name	Color
annotate1	Green3	Bright Green
annotate2	Deeppink	Bright Pink
annotate3	Cyan1	Bright Blue
annotate4	Gold1	Orange
annotate5	Chocolate1	Brown
annotate6	Red1	Red
annotate7	Darkorchid1	Dark Purple
annotate8	Wheat3	Yellow-Orange
annotate9	Blue3	Normal Blue
annotate10	Darkolivegreen1	Dark Green
annotate11	Khaki4	Yellow
annotate12	Lightpink2	Light Pink
annotate13	Dodgerblue4	Deep Blue

2.3.1.2 Determining significant SNPs

The next step was to select the significant SNPs associated with each phenotype. The Perl script called "FindingSignificantSNPs.pl" (Appendix 3) reads each .ps file and saves the SNPs with a *p*-value equal to or less than the significant cutoff determined by the user into a new file with the ending ".sigsnps.csv".

The "FindingSignificantSNPs.pl" script has two different parts to it. The first part opens and reads the file containing the allele frequencies of all the SNPs. The second part reads the .ps file. The allele frequency of the SNP is then added to the information saved for it. The output prints into a file named "phenotype.sigsnp.csv" and is saved into a folder called "SignificantSNPs". After the script has finished the folder "SignificanSNPs" should contain a file for each phenotype that has the following column names: Trait, SNP, CHR, BP, P, Beta, and Non-Col_AlleleFreq (see Table 2.2).

To decrease the required running time the SNP order from the allele frequency file must match the SNP order of the genotype. This is only a concern when using the SNP files with a criterion MAF ≤ 5%. To correct for this problem, an additional command was added to the Perl script to eliminate the SNPs from the allele frequency file as the file was read. SNPs with MAF ≤ 5% were not saved in the array of allele frequencies. This ensures that the SNP order of the saved allele frequencies will be the same order as the SNP order of the genotype file, and decrease the time required to run this script.

Table 2.2. Example of the significant SNP output. Columns are the following: Trait is the phenotype, SNP is the SNP ID, CHR is the chromosome, BP is the base pair position, P is the *p*-value, Beta is the β-value, and Non-Col_AlleleFreq is the frequency of the non-Columbia allele.

Trait	SNP	CHR	BP	P	Beta	Non-Col_ AlleleFreq
GrandMean	15923981	1	5923981	7.20E-05	0.333600105	0.044392523
GrandMean	15924484	1	5924484	6.08E-05	0.439481774	0.028037383
GrandMean	15924553	1	5924553	8.15E-05	0.415410352	0.030373832
GrandMean	15925216	1	5925216	6.08E-05	0.439481774	0.028037383
GrandMean	15925576	1	5925576	6.54E-05	0.420568443	0.030373832
GrandMean	15925820	1	5925820	6.08E-05	0.439481774	0.028037383
GrandMean	15926015	1	5926015	6.08E-05	0.40882258	0.03271028
GrandMean	19614666	1	9614666	6.56E-05	-0.151669099	0.528037383

2.3.1.3 Calculating the false discovery rate

The standard significant cutoff (p-value < $1x10^{-5}$) is less than a Bonferroni corrected cutoff (p-value < $2.37x10^{-7}$) for 211K SNPs. To try to eliminate false positive SNPs, false discovery rate (FDR) was used to recalculate the significance of each SNP. Another Perl script, called "CalculateFDR.pl" read each .ps file and calculated the significance of each SNP based on the Benjamini-Hochberg theory, as explained above. The output was printed in a new file called "phenotype.fdr.csv" and saved in a folder called "FDR_SignificantSNPs." The folder contained a file for each phenotype, and each file contained 10 columns containing the new p-value calculated using the Benjamini-Hochberg theory (Table 2.3).

2.3.2 Running MLMM

A second pipeline was created to create putative genes lists from significant SNPs calculated using MLMM (Segura et al. 2012). The MLMM analysis runs in R, and the output includes manhattan plots, p-values of SNPs as determined by most significant BIC and Bonferroni models, and the significant SNPs for each model. A cutoff of the p-values limits the number of SNPs saved in the p-value files. The user determined the cutoff for which p-values are printed. The R script originally printed only two files of significant SNPs, one containing the significant SNPs as determined the BIC model and the second containing the significant SNPs as determined by the Bonferroni model. Modifications to the R script allowed the user to see what the significant SNPs were for each iteration of the statistical model. The number of iterations was determined by the user by determining

Table 2.3. Example of the FDR significant SNP output. Columns are the following: Trait = phenotype, SNP = SNP ID, CHR = chromosome, BP = base pair position, P = *p*-value, Beta = β-value, Non-Col_AlleleFreq = the frequency of the non-Columbia allele, FDR-derivedSignificantThreshold = the new significant cutoff, FDR-adjected P-value = the new *p*-value of the SNP, and Rank = the order of SNPs based on initial *p*-value.

Trait	SNP	CHR	BP	P	Beta	Non-Col_ AlleleFreq	FDR-derived Significant Threshold	FDR-adjusted P-value	Rank
0.25X_0.5X	212002218	2	12002218	8.56E-07	0.216516716	0.439252336	2.06E-06	0.020815841	7
0.25X_0.5X	212006389	2	12006389	4.99E-08	-0.392966762	0.892523364	2.94E-07	0.008494128	1
0.25X_0.5X	212006825	2	12006825	7.71E-08	-0.376122585	0.885514019	8.81E-07	0.004374731	3
0.25X_0.5X	212007650	2	12007650	6.26E-08	-0.374773481	0.880841121	5.87E-07	0.00532798	2
0.25X_0.5X	212008374	2	12008374	1.59E-07	-0.362987305	0.88317757	1.17E-06	0.006766364	4
0.25X_0.5X	212008536	2	12008536	4.81E-07	-0.333300666	0.869158879	1.76E-06	0.013646211	6
0.25X_0.5X	212009943	2	12009943	2.04E-07	-0.360049527	0.88317757	1.47E-06	0.006945098	5

the number of maximum steps that should be calculated. If the number of maximum steps were set to 10 then 9 significant SNPs would be added to the model by the end of the script. The top SNPs determined by the BIC and Bonferroni models were always included in the top 9 SNPs, but were usually a different model, with fewer SNPs included. However, by saving the top SNP selected from each iteration the user has more information to parse through to find genes affecting their phenotype.

The significant SNPs are printed into a file as the SNP ID only. We wrote a Perl script that opens all the significant SNP files and parses the SNP ID into the chromosome and base pair position, and finds the *p*-value of the each SNP in the *p*-value output file and prints off a new significant SNP file for each phenotype and model. Each file has the following columns: Trait, SNP, CHR, BP, P (See Table 2.4). Modifying the significant SNP file is done to match the significant SNP files created from EMMAX output.

MLMM does not allow for missing data; therefore, a new genotype has to be created for each group of phenotypes so that all accessions included in the model does have data. However, a new allele frequency file has to be created for each new set of phenotypes. As of now, the allele frequency is not included in the output of MLMM, but it can easily be done by creating new allele frequency files.

2.3.3 Defining putative gene lists

Once the significant SNPs are determined and put into the correct file format for the EMMAX and MLMM output, another Perl script is used to link genes to the SNPs. The script, called "DetermineGenes_LinkedToSigSNPs.pl", opens and reads the significant SNP files. The script is saved into the same folder as the files containing the SNPs.

Table 2.4. Example of the significant SNP file created from MLMM output. The name of the trait contains which model the significant SNPs were found. Columns are the following: Trait = phenotype, SNP = SNP ID, CHR = chromosome, BP = base pair position, and P = p-value.

Trait	SNP	CHR	BP	P
GrandMean_BICcof.csv	518566007	5	18566007	1.11E-09
GrandMean_BICcof.csv	521571544	5	21571544	8.55E-11
GrandMean_BICcof.csv	210697985	2	10697985	2.58E-12
GrandMean_BICcof.csv	41328437	4	1328437	3.35E-08
GrandMean_BICcof.csv	127295432	1	27295432	3.02E-07

The script has three different parts: Reading the gene description file, the gene information file, and finding the closest gene to each significant SNP and the five genes upstream and five genes downstream of the significant SNP. Each SNP is linked to 11 genes. The information is printed into a new file named "phenotype.candigenes.csv" and saved in new folder named "CandidateGenes."

The first part of the script opens and reads a file called "TAIR10_functional_descriptions." This file was downloaded from arabidopsis.org and contains the model name of genes, e.g. AT1G01010.1, the gene description, and the function of the gene product. The script read the file and parsed the information into different columns. The model name was modified so that it was only the gene name (AT1G01010). The different gene models were eliminated and only one copy of each gene was kept under the gene name (AT1G01010). Only one gene description was saved with each gene.

The second part of the script opened and read the file called "TAIR10_GFF3_genes.gff", which contained the mapping information of the genes. This file was also downloaded from arabidopsis.org. This file was parsed into different columns. The mitochondrial and chloroplastic genes were eliminated from the list of genes, and only genes with the label of "gene" were saved. The gene description information and the mapping information were matched up, and pushed onto an array, so that each gene had its mapping position and gene function description. The genes were then sorted according to the start position of the genes.

The significant SNP files were then opened one at a time and using the SNP base pair position, the gene that contained the SNP or that was closest to the SNP was labeled

as the “hit” gene. The five genes upstream and downstream of the hit gene were labeled “Up1, Up2, Up3…down3, down2, down1”. These eleven genes for each significant SNP were printed into the new candidate gene file. The information for each SNP and gene include the SNP ID, SNP base pair position, the chromosome, the gene position (hit or up/down), gene model name, and gene descriptions (see Table 2.5).

2.4 Conclusion

The pipeline for running GWA using two different methods was created to decrease the time required to run several phenotypes at a time by eliminating the need to execute every phenotype as a separate entity. By using the pipeline, the user can submit a file containing as many phenotypes as wanted and the final output are putative gene lists for each phenotype of genes that may be contributing to the trait.

Each method has pros and cons associated with each. Both EMMAX and MLMM are able to compute small SNP data sets (211K SNPs) and give results that are manageable. However, with large data sets, 4.9 million SNPs, each method has its downfalls.

For large SNP datasets EMMAX was faster than MLMM because it calculated the associations fairly quickly. However, the number of significant SNPs is too high to be true. Also, for some traits, the associations calculated for SNPs with a low allele frequency (MAF $\leq$ 5%) were equal indicating that thousands of SNPs through the genome were significant. By eliminating these SNPs from the genotype file, these associations disappeared, but most often the number of significant SNPs still remained

Table 2.5. Example of the putative gene lists created for each phenotype. Columns are the following: Trait = phenotype, SNP = SNP ID, B_Value = β-value, P = *p*-value, Non_Col_Allele_freq = non-Columbia allele frequency, Position = gene location of SNP, Type = gene, Chr = chromosome, Start_bp = start position of gene, End_bp = end position of gene, Gene = gene locus, Name = Description. Two other columns are also included, one is another gene description and the second contains the hyperlink to the gene on arabidopsis.org.

Trait	SNP	B_value	P-value	Non_Col_ Allele_freq	Position	Type	Chr	Start_bp	End_bp	Gene	Name
GrandMean	15923981	0.333600105	7.20E-05	0.044392523	Hit	gene	1	5922630	5926400	AT1G17290	alanine aminotransferas (AlaAT1)
GrandMean	15923981	0.333600105	7.20E-05	0.044392523	Down5	gene	1	5901169	5903439	AT1G17250	receptor like protein 3 (RLP3)
GrandMean	15923981	0.333600105	7.20E-05	0.044392523	Down4	gene	1	5904058	5908898	AT1G17260	autoinhibited H(+)-ATPase isoform 10 (AHA10)
GrandMean	15923981	0.333600105	7.20E-05	0.044392523	Down3	gene	1	5909249	5911693	AT1G17270	O-fucosyltransferase family protein
GrandMean	15923981	0.333600105	7.20E-05	0.044392523	Down2	gene	1	5916871	5920089	AT1G17280	ubiquitin-conjugating enzyme 34 (UBC34)
GrandMean	15923981	0.333600105	7.20E-05	0.044392523	Down1	gene	1	5920774	5921415	AT1G17285	
GrandMean	15923981	0.333600105	7.20E-05	0.044392523	Up1	gene	1	5926994	5927784	AT1G17300	
GrandMean	15923981	0.333600105	7.20E-05	0.044392523	Up2	gene	1	5928014	5928667	AT1G17310	MADS-box transcription factor family protein
GrandMean	15923981	0.333600105	7.20E-05	0.044392523	Up3	gene	1	5929909	5931831	AT1G17330	Metal-dependent phosphohydrolase
GrandMean	15923981	0.333600105	7.20E-05	0.044392523	Up4	gene	1	5933870	5938761	AT1G17340	Phosphoinositide phosphatase family protein
GrandMean	15923981	0.333600105	7.20E-05	0.044392523	Up5	gene	1	5940423	5941210	AT1G17345	SAUR-like auxin-responsive protein family

high. Having a high number of significant SNPs makes interpretation of the data more difficult because the user has to decide which SNPs and genes to follow.

The main problem with MLMM is that it is executed in R. The memory capacity of R limits the number of SNPs MLMM can calculate. Even loading the SNP file into R would take 10 -20 hours and usually resulted in the software freezing up. For 211K SNPs, each iteration takes several minutes; therefore, depending on the number of iterations to calculate the associations of one phenotype could take hours. For 1.6M SNPs, each iteration takes a couple of hours. Whereas, EMMAX only takes 5 minutes to complete one analysis using 1.6 million SNPs, MLMM takes hours. Ideally, the benefit of MLMM is that 4.9 million SNPs could be used to calculate associations because of the multi-locus testing should eliminate all the spurious positive hits and the SNPs with low allele frequencies should not have the same problem as they did with EMMAX. However, the computational power to execute 4.9M SNPs using MLMM was too high, and the script failed.

Another benefit of MLMM is the amount of significant SNPs given in the output. MLMM uses stringent model selections to find the best model that fits the phenotypic data. Only a small number of SNPs are significant per phenotype. Potentially, interpretation is a lot easier and deciding which genes to do follow-up experiments would be less subjective. The output of MLMM potentially requires less a priori knowledge and hopefully the putative genes linked to the significant SNPs would be potential candidate genes for the desired trait.

In conclusion, both methods of calculating associations have their benefits and their negative aspects. Each method uses a linear mixed model and the software, EMMA,

with modifications to execute their statistical models. Depending on the users purpose one method could be used over the other.

2.5 References

Abdollahi-Arpanahi, R, a Pakdel, a Nejati-Javaremi, M Moradi Shahrbabak, G Morota, BD Valente, a Kranis, GJM Rosa, D Gianola (2014) Dissection of additive genetic variability for quantitative traits in chickens using SNP markers. J Anim Breed Genet 131:183–93

Atwell, S, YS Huang, BJ Vilhjálmsson, G Willems, M Horton, Y Li, D Meng, A Platt, AM Tarone, TT Hu, R Jiang, NW Muliyati, X Zhang, MA Amer, I Baxter, B Brachi, J Chory, C Dean, M Debieu, J de Meaux, JR Ecker, N Faure, JM Kniskern, JDG Jones, T Michael, A Nemri, F Roux, DE Salt, C Tang, M Todesco, MB Traw, D Weigel, P Marjoram, JO Borevitz, J Bergelson, M Nordborg (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627–31

Benjamini, Y, Y Hochberg (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B (Statistical Methodol 57:289–300

Benjamini, Y, Y Hochberg (2000) On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics. J Educ Behav Stat 25:60–83

Benjamini, Y, AM Krieger, D Yekutieli (2006) Linear Step-up Procedures That COntraol the False Discovery Rate. Biometrika 93:491–507

Cao, J, K Schneeberger, S Ossowski, T Günther, S Bender, J Fitz, D Koenig, C Lanz, O Stegle, C Lippert, X Wang, F Ott, J Müller, C Alonso-Blanco, K Borgwardt, KJ Schmid, D Weigel (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43:956–63

Chan, EKF, HC Rowe, DJ Kliebenstein (2010) Understanding the Evolution of Defense Metabolites in Arabidopsis thaliana Using Genome-wide Association Mapping. Genetics 185:991–1007

Cho, YS, MJ Go, YJ Kim, JY Heo, JH Oh, H-J Ban, D Yoon, MH Lee, D-J Kim, M Park, S-H Cha, J-W Kim, B-G Han, H Min, Y Ahn, MS Park, HR Han, H-Y Jang, EY Cho, J-E Lee, NH Cho, C Shin, T Park, JW Park, J-K Lee, L Cardon, G Clarke, MI McCarthy, J-Y Lee, J-K Lee, B Oh, H-L Kim (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat Genet 41:527–34

Cupples, LA, HT Arruda, EJ Benjamin, RB D'Agostino, S Demissie, AL DeStefano, J Dupuis, KM Falls, CS Fox, DJ Gottlieb, DR Govindaraju, C-Y Guo, NL Heard-Costa, S-J Hwang, S Kathiresan, DP Kiel, JM Laramie, MG Larson, D Levy, C-Y Liu, KL Lunetta, MD Mailman, AK Manning, JB Meigs, JM Murabito, C Newton-Cheh, GT O'Connor, CJ O'Donnell, M Pandey, S Seshadri, RS Vasan, ZY Wang, JB Wilk, P a Wolf, Q Yang, LD Atwood (2007) The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. BMC Med Genet 8 Suppl 1:S1

Devlin, B, K Roeder (1999) Genomic Control for Association Studies. Biometrics 55:997–1004

Florez, JC, AK Manning, J Mcateer, K Irenze, L Gianniny, DB Mirel, CS Fox, LA Cupples, JB Meigs (2007) A 100K Genome-Wide Association Scan for Diabetes and Related Traits in the Framingham Heart Study. Diabetes 56:3063–3074

Freudenberg, J, H-S Lee, B-G Han, H Do Shin, YM Kang, Y-K Sung, S-C Shim, C-B Choi, AT Lee, PK Gregersen, S-C Bae (2011) Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci. Arthritis Rheum 63:884–93

Horton, MW, AM Hancock, YS Huang, C Toomajian, S Atwell, A Auton, NW Muliyati, A Platt, FG Sperone, BJ Vilhjálmsson, M Nordborg, JO Borevitz, J Bergelson (2012) Genome-wid patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel 44:212–216

Kang, HM, JH Sul, SK Service, N a Zaitlen, S-Y Kong, NB Freimer, C Sabatti, E Eskin (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348–54

Kang, HM, N a Zaitlen, CM Wade, A Kirby, D Heckerman, MJ Daly, E Eskin (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–23

Li, Y, Y Huang, J Bergelson, M Nordborg, JO Borevitz (2010) Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. Proc Natl Acad Sci U S A 107:21199–204

Malosetti, M, CG van der Linden, B Vosman, F a van Eeuwijk (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. Genetics 175:879–89

Patterson, N, AL Price, D Reich (2006) Population structure and eigenanalysis. PLoS Genet 2:e190

Pike, N (2011) Using false discovery rates for multiple comparisons in ecology and evolution. Methods Ecol Evol 2:278–282

Platt, A, M Horton, YS Huang, Y Li, AE Anastasio, NW Mulyati, J Agren, O Bossdorf, D Byers, K Donohue, M Dunning, EB Holub, A Hudson, V Le Corre, O Loudet, F Roux, N Warthmann, D Weigel, L Rivero, R Scholl, M Nordborg, J Bergelson, JO Borevitz (2010) The scale of population structure in Arabidopsis thaliana. PLoS Genet 6:e1000843

Price, AL, NJ Patterson, RM Plenge, ME Weinblatt, N a Shadick, D Reich (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–9

Pritchard, JK, M Stephens, NA Rosenberg, P Donnelly (2000) Association mapping in structured populations. Am J Hum Genet 67:170–81

R Core Team (2013) R: A language and environment for statistical computing. Vienna, Austria

Sabatti, C, SK Service, A-L Hartikainen, A Pouta, S Ripatti, J Brodsky, CG Jones, N a Zaitlen, T Varilo, M Kaakinen, U Sovio, A Ruokonen, J Laitinen, E Jakkula, L Coin, C Hoggart, A Collins, H Turunen, S Gabriel, P Elliot, MI McCarthy, MJ Daly, M-R Järvelin, NB Freimer, L Peltonen (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 41:35–46

Segura, V, BJ Vilhjálmsson, A Platt, A Korte, Ü Seren, Q Long, M Nordborg (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44:825–30

Storey, JD (2002) A direct approach to false discovery rates. J R Stat Soc Ser B 64:479–498

Tabangin, ME, JG Woo, LJ Martin (2009) The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proc 3 Suppl 7:S41

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–78

Tian, F, PJ Bradbury, PJ Brown, H Hung, Q Sun, S Flint-Garcia, TR Rocheford, MD McMullen, JB Holland, ES Buckler (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet 43:159–62

Todd, J a, NM Walker, JD Cooper, DJ Smyth, K Downes, V Plagnol, R Bailey, S Nejentsev, SF Field, F Payne, CE Lowe, JS Szeszko, JP Hafler, L Zeitels, JHM Yang, A Vella, S Nutland, HE Stevens, H Schuilenburg, G Coleman, M Maisuria, W Meadows, LJ Smink, B Healy, OS Burren, A a C Lam, NR Ovington, J Allen, E Adlem, H-T Leung, C Wallace, JMM Howson, C Guja, C Ionescu-Tîrgovişte, MJ Simmonds, JM Heward, SCL Gough, DB Dunger, LS Wicker, DG Clayton (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39:857–64

Turner, S (2011) GettingGeneticsDone. http://gettinggeneticsdone.blogspot.com/

Verhoeven, KJF, KL Simonsen, LM Mcintyre (2005) Implementing false discovery rate control : increasing your power. OIKOS 108:643–647

Wray, NR, SM Purcell, PM Visscher (2011) Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol 9:e1000579

Yu, J, G Pressoir, WH Briggs, I Vroh Bi, M Yamasaki, JF Doebley, MD McMullen, BS Gaut, DM Nielsen, JB Holland, S Kresovich, ES Buckler (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–8

Zeggini, E, MN Weedon, CM Lindgren, TM Frayling, KS Elliott, H Lango, NJ Timpson, JRB Perry, NW Rayner, RM Freathy, JC Barrett, B Shields, AP Morris, S Ellard, CJ Groves, LW Harries, JL Marchini, KR Owen, LR Cardon, M Walker, GA Hitman, AD Morris, JC Florez, H Chen, J Meyer, N Joel, C Study, S Kathiresan, T Leader, MJ Daly, TE Hughes, C Guiducci, A Surt, NP Burtt, O Melander, P Almgren, HN Lyon, Q Ma, H Parikh, D Richardson, D Ricke, J Roix, L Groop, S Purcell, C Newton-cheh, P Nilsson, M Orho, M Daly-, D Altshuler, JN Hirschhorn, R Tewhey, R Barry, W Brodeur, P Burtt, J Camarata, N Chia, MF John, S Elliott, P Morris, R Owen (2007) Replication of Genome-Wide Association Reveals Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes. Science (80-) 316:1336–1341

Zhao, K, MJ Aranzana, S Kim, C Lister, C Shindo, C Tang, C Toomajian, H Zheng, C Dean, P Marjoram, M Nordborg (2007) An Arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4

CHAPTER 3. PIPELINE LINKING GENOTYPE TO PHENOTYPE USING GENOME-WIDE ASSOCIATION

3.1 Introduction

The natural variation of *A. thaliana* has been well studied because of its ideal speciation patterns (Koornneef et al. 2004, Platt et al. 2010, Fournier-Level et al. 2011, Horton et al. 2012, Weigel 2012). The species has spread from Northern Europe to Southern Europe and has diverged immensely. It has a short life span, and has a high reproduction rate in the lab. Also, *A. thaliana* is a selfing plant, resulting in individuals that are highly homozygous, and a population that is highly polymorphic. This is ideal for genome-wide association (GWA) studies, which uses SNPs to map genotype to phenotypes at smaller genetic windows than other mapping tools such as QTL. Over the past few years, a number of GWA studies has come out using *A. thaliana* to map genes to phenotypes (Chan et al. 2010, Atwell et al. 2010, Li et al. 2010, Filiault and Maloof 2012, Horton et al. 2014).

However, using *A. thaliana* for GWA does have disadvantages. Atwell et al. (2010) found that the mixed-model did perform well, but results were still biased because of the population structure of *A. thaliana*. They concluded that knowing which associations were true, and which were worth furthering study was highly subjective and a priori knowledge of the trait was highly helpful (Atwell et al. 2010). Therefore, interpretation of GWA results were not easily made and further study of the GWA results

required careful consideration of all positive associations to determine which ones would be used in follow-up studies.

In this chapter, genomic associations were calculated using GWA for four different phenotypic datasets. Using the pipeline described in Chapter 2 to run the multi-phenotypic datasets, determine significant SNPs, and to create putative gene lists from EMMAX and MLMM results. The phenotypes that were studied were glufosinate tolerance, hybrid incompatibility, seed size, and secondary metabolites.

3.1.1 Glufosinate tolerance

Glufosinate is a common foliar herbicide. Genetically modified crops, such as soybean, canola, cotton, and maize, are resistant to glufosinate. These crops are being introduced and help minimize weed management; however, weeds are also evolving resistance to glufosinate (Heap 2015). Herbicide tolerance can evolve within the target-site gene (target-site resistance) or at a non-target site genes (non-target-site resistance) (Gardner et al. 1998, Yuan et al. 2007, Yu et al. 2009). Target-site resistance is more easily mapped than non-target-site resistance because the gene of function is known. Non-target-site resistance may involve one locus or multiple loci, increasing the difficulty of learning the genetic mechanism of non-target-site resistance. Discovering the mechanism of non-target site resistance requires understanding the genetic, biological, and biochemical effects of the plant responses to the herbicide.

Glufosinate competes with glutamine for the binding site of glutamine synthetase (GS) (Manderscheid and Wild 1986, Lacuesta et al. 1990). GS is required for the reassimilation of ammonium released during photorespiration and other biological processes by binding glutamate and ammonium to synthesize glutamine (Keys et al. 1978,

Miflin and Lea 1980, Robertson and Farnden 1980, Woo et al. 1982, Wallsgrove et al. 1983). Six paralogs of *GLUTATMINE SYNTHETASE* (*GS*) are found in the *A. thaliana* genome, and only *GLUTAMINE SYNTHETASE2* (*GS2*) encodes the GS that specifically reassimilates the ammonium released during photorespiration.

A popular hypothesis for the glufosinate mode of action is that ammonium accumulates at toxic levels and inhibits photosynthesis (Abdeen and Miki 2009, Seabra et al. 2012). However, it has been shown that plants survived glufosinate toxicity when placed in an atmosphere that inhibits photorespiration (Morris et al. 1989, Wendler et al. 1990). Wendler et al. (1990) also showed that glufosinate toxicity was reduced by treating plants with specific amino acids (Wendler et al. 1990). These plants accumulated high levels of ammonium still, but the additional amino acids allowed photosynthesis to continue to function (Wendler et al. 1990). This would suggest that high levels of ammonium is not the cause of toxicity, but the lack of amino acid biosynthesis, because ammonium reassimilation is inhibited, lead to failure of other biological processes such as photosynthesis and eventual cell death.

Photorespiration is a complex pathway that requires the regeneration of the Calvin Cycle substrate 3-phosphoglycerate from 2-phosphoglycolate (Givan et al. 1988, Peterhansel et al. 2010). Photorespiration releases CO_2 and ammonium as byproducts that need to be reassimilated (Keys et al. 1978, Peterhansel et al. 2010). The ammonium is reassimilated by GS into glutamine, which acts as a building block for amino acid synthesis (Woo et al. 1982, Walker et al. 1984). Amino acid content changes when photorespiration is disrupted suggesting that photorespiration is critical for amino acid production (Somerville and Ogren 1983, Walker et al. 1984, Wallsgrove et al. 1986,

Wendler et al. 1990, Häusler et al. 1994, Leegood et al. 1995, Novitskaya et al. 2002). Though the primary amino acids involved in photorespiration are glycine, serine, glutamine, and glutamate other amino acids are used as substrates in photorespiration (Ta and Joy 1986, Givan et al. 1988, Bauwe et al. 2010). The complex photorespiration pathway involves several essential biological pathways. It is possible that glufosinate tolerance results from changes in these biological pathways that compensate for GS2 inhibition (Potel et al. 2009).

The physiological changes that occur after glufosinate application are broad and change with time. Plant response to glufosinate has an early and a late stage (Abdeen and Miki 2009). Abdeen and Miki (2009) showed that genes involved in amino acid metabolism, secondary metabolism, transcription, hormonal regulation, plant defense response, detoxification, cell death, photosynthesis, and developmental process change transcript concentrations upon glufosinate application (Abdeen and Miki 2009). We hypothesized that variation within these affected biological processes can lead to non-target-site resistance (Yuan et al. 2007, Yu et al. 2009, Kaundun 2010). Selection for glufosinate-resistant polymorphisms within the *GS* genes may not occur, but selection for glufosinate-resistant polymorphisms in one or more genes of one or more biological process could occur resulting in non-target-site resistance.

3.1.2 Hybrid incompatibility

A. thaliana has been a model species for studying the genetics of hybridization barriers, in both interspecies and interpolyploidy hybrids (Adams et al. 2000, Comai et al. 2000, Henry et al. 2005, 2007, Josefsson et al. 2006, Dilkes et al. 2008, Walia et al. 2009, Chang et al. 2010, Burkart-Waco et al. 2012, 2013, Kradolfer et al. 2013b, Schatlowski et

al. 2014). The two types of hybrids show very similar seed phenotype, which is high seed abortion during early embryogenesis (Haig and Westoby 1991, Scott et al. 1998, Bushell et al. 2003, Dilkes et al. 2008, Burkart-Waco et al. 2012). Seed abortion is the result of the misregulation of endosperm development, which then leads to embryo abortion (Scott et al. 1998, Dilkes et al. 2008, Hehenberger et al. 2012, Kradolfer et al. 2013a). Many studies have focused on understanding which genes contribute to the regulation of endosperm development, and thus to hybridization barriers (see above references).

Imprinted genes play a critical role in endosperm development (Scott et al. 1998, Kradolfer et al. 2013a, 2013b, Schatlowski et al. 2014). One example is the Polycomb-group (PcG) proteins that form a complex called the Polycomb-like Repressive Complex2 (PRC2). The PRC2 is required for repressing the development of the endosperm from the central cell within the ovule before fertilization (Chaudhury et al. 1997, Köhler et al. 2003). When PRC2 is misregulated then the endosperm starts to develop before fertilization. Upon fertilization, these seeds begin to develop, but then abort at early embryogenesis, similar to the seed abortion in hybrid crosses, suggesting that the PRC2 is involved the hybridization barrier. In fact, it was found that the deregulation of the *MEDEA* (*MEA*) gene, part of PRC2, complements the interploidy seed phenotype, and more seeds successfully reached maturity when *MEA* was deregulated (Erilova et al. 2009).

The Köhler lab has been extensively studying the effect of imprinting on interploidy hybridizations (Hehenberger et al. 2012, Kradolfer et al. 2013a, 2013b, Schatlowski et al. 2014). Schatlowski et al. (2014) found that by decreasing methylation in the pollen the triploid block was passed, and viable seed was produced in an

interploidy hybridization (Schatlowski et al. 2014). They suggested that hypomethylation leads to new CHG methylation patterns of the PRC2 target genes. The PRC2 complex regulates gene expression using methylation. When the PRC2 complex is inhibited seed death occurs showing that gene regulation by PRC2 is critical for seed development (Chaudhury et al. 1997, Guitton and Berger 2005). However, changing the methylation patterning of PRC2 target genes overcomes the effects of an inhibited complex, suggesting that the de novo methylation pattern created upon pollination of hypomethylation pollen is sufficient genetic control for seed development (Schatlowski et al. 2014).

Supporting the hypothesis that hybridization barriers are created by deregulated methylation and thus deregulation of imprinted genes in the developing seed means that the parental genomes are differentially contributing genetic information to the seed. This is reasonable in sight of the parental conflict. The maternal plant expresses genes that control resources so all fertilized ovules receive equal proportions, and the paternal pollen expresses genes that draw nutrients into that specific fertilized ovule (Haig and Westoby 1989). Many studies concluded that most genes are differently expressed in the early zygote (Vielle-Calzada et al. 2000, Baroux et al. 2001, 2008, Grimanelli et al. 2005, Autran et al. 2011). However, differentiating between genes expressed in the sporophyte verses the zygote is difficult because the tissues are microscopic and contamination occurs easily.

One study isolated the zygote at the 1-cell, 2-cell, 8-cell, and 32-cell stages and measured which parental transcriptome contributed most during early zygotic development (Nodine and Bartel 2012). They found that both parental transcriptomes contributed evenly during these stages of embryogenesis; therefore, the models suggesting that most gene

expression comes from the maternal genome were false. They also hypothesized that the rapid turnover of gene expression in the zygote overrode any effects of the different RNA contributions made by either parents (Nodine and Bartel 2012). They did find 122 genes that showed a parental bias at one or two of the different stages (Nodine and Bartel 2012).

Another factor that plays a role in hybridization barriers is the role of the sporophytic tissue of the ovule—the development of the seed coat (Dilkes et al. 2008). Seed development relies upon a careful control of the growth of the endosperm, embryo, and seed coat (Garcia et al. 2003, 2005, Luo et al. 2005, Berger et al. 2006, Nowack et al. 2010, Hehenberger et al. 2012). Dilkes et al. (2008) found that by inhibiting the accumulation of proanthocyanidins, a type of flavonoid, in the integuments that interploidy seed development appeared more normal than in wildtype interploidy crosses and the rate of lethality decreased in the hybrid crosses (Dilkes et al. 2008). This suggested a role of flavonoid production in the hybridization barriers, a possible hypothesis being that resources needed for anthocyanin and proanthocyanidin synthesis can be reallocated for the use of the developing endosperm and embryo.

3.1.3 Seed size

Seed size is regulated by endosperm, embryo, and seed coat development (Garcia et al. 2003, 2005). The endosperm must develop correctly, as the endosperm is the source of nutrients to the embryo. The seed coat must differentiate and grow to create space for the growing endosperm and embryo. Larger endosperms can give more nutrients to the growing embryo, but those resources are costly. Seed size is determined by balancing size verses number of seed. Large seeds are more costly than small seeds, limiting the number of seeds the plant can produce (Smith and Fretwell 1974, Haig and Westoby

1991). The parental conflict suggests that the maternal genetic contributions restrict seed growth and the paternal genetic contributions promote seed growth. In support of this, it has been shown that hybrid seeds that had an excess of maternal genome were smaller verses hybrid seed with an excess of paternal genome, which were larger than wildtype (Scott et al. 1998). Since seed size is affected by the parental genetic contributions and is controlled by endosperm development, it has been hypothesized that seed size contributes to hybridization barriers (Haig and Westoby 1991).

In this thesis, we hypothesized that the genetic mechanisms that control seed size should also contribute to hybrid incompatibilities. The GWA results from both hybrid incompatibilities and seed size should share common significant SNPs for genes that are contributing to both traits.

3.1.4 Secondary metabolites

Trying to understand the evolutionary consequences and the genetic mechanisms of producing so many diverse secondary metabolites is widespread (Hartmann 2007). With more than 200,000 different metabolites synthesized, understanding the biosynthesis pathways have not been easy (Dixon and Strack 2003, Hartmann 2007, Yonekura-Sakakibara and Saito 2009). Even so, through genetic research, the pathways for metabolites such as flavonoids, sinapate esters, lignin, terpenoids, and glucosinolates have been elucidated (Shirley et al. 1995, Kliebenstein et al. 2001a, 2001b, Kroymann et al. 2001, Ruegger and Chapple 2001, Chen et al. 2003, Tholl et al. 2005). Despite all the research dedicated to metabolite synthesis, much of secondary metabolite biosynthesis and function remains unknown (D'Auria and Gershenzon 2005).

A recent study demonstrated the usefulness of GWA in elucidating the biosynthesis pathway of secondary metabolites (Li et al. 2014). Taking advantage of the natural variation of *A thaliana* in both genetics and in secondary metabolites, Li et al. (2014) uncovered a UDP glycosyltransferase gene that is required for the synthesis of four different dihydroxybenzoic acid glycosides (Li et al. 2014). Using GWA to link genotype to phenotype is a fast and efficient tool for initially looking into studying any secondary metabolite.

3.1.5 Summary

These four different traits are complex and multigenic. Understanding the genetic mechanism for these four different traits is challenging and complex. Through GWA, insights can found by mapping genotypes to phenotype using *A. thaliana* as a model because of the high natural variation that exist within this species. My pipeline, used to calculate GWA for multiple phenotypes using two different statistical methods, EMMAX and MLMM, was fast and efficient. My pipeline created putative gene lists for each trait. The results for all four different traits studied are available for the general public, and can be used to find new candidate genes contributing to each of these phenotypes. We summarized some of our findings in the remainder of the chapter.

3.2 Glufosinate tolerance

Although A. thaliana has not naturally been under selection for glufosinate tolerance, we hypothesized that the natural variation in biological processes found in *A. thaliana* would be sufficient to find candidate non-target-site resistant genes contributing to glufosinate tolerance using GWA.

3.2.1 Methods

3.2.1.1 Plant material

For the GWA, the 440 accessions used in the study consisted of the 360 accessions used for scaling *A. thaliana* population structure and the first 80 accessions sequenced (Platt et al. 2010, Cao et al. 2011). Not all 440 accessions grew neither was SNP data available for all the accessions. Thus, 428 accessions were used in the 211K SNPs dataset and 430 accessions were used in the 1.6M SNPs dataset.

For testing candidate genes the following SALK lines were ordered from TAIR (arabidopsis.org). *hsi2* (AT2G30470)—SALK_088606C; *ivd1* (AT3G45300)—CS860822; *gpat8* (AT4G00400)—SALK_043084C; *aop3* (AT4G03050)—SALK_001655C; *aop1* (AT4G03070)—SALK_06735; *shm4* (AT4G13930)—SALK_054155C; *shm3* (AT4G32520)—SALK_113687C; *shm1* (At4g37930)—SALK_089133C; K23L20.6 (AT5G44720)—SALK_081127C; *spds3* (AT5G53120)—SALK_018902C.

3.2.1.2 GWA planting and phenotyping design

Two sets of the 440 accessions were planted. The seeds were randomly planted on 32-cell trays, and stratified for one week. The plants were then placed in a growth room. One set was sprayed with 0.25X glufosinate mixed with an adjuvant. The other set was sprayed with 0.125X glufosinate mixed with an adjuvant. The plants were scored 6 days after spraying (DAS) and 14 DAS. A 1-5 scale was used to score glufosinate damage. The scale was: 1= healthy, no damage to leaves; 2= 2-3 leaves showed damage; 3= 4-5 leaves showed damage; 4= all leaves damaged and only meristemic tissue is alive; 5=

dead. After 14 days both sets were sprayed with 0.5X glufosinate mixed with an adjuvant, and the damage scored again. Along with the basic scores, different averages were calculated to try and accurately measure glufosinate tolerance. Thus, a total of 10 phenotypes were measured and calculated: 0.25X_6DAS, 0.25X_14DAS, 0.25X_0.5X, 0.125X_6DAS, 0.125_14DAS, 0.125X_0.5X, Grand Mean (the average of all 6 scores for each accession), Mean_14DAS (the average score for each accession using the two 14 DAS scores from each spray), Mean_6DAS (the average score for each accession using the two 6 DAS scores), and Mean0.5X (the mean for the score of each accession after being sprayed with 0.5X glufosinate from the two different sets).

Bonferroni corrections were calculated to correct for multiple testing. The Bonferroni method simply divides the significant cutoff by the number of hypotheses tested (see chapter 2 for more explanation).

3.2.1.3 Testing candidate genes

10 replicates of knock-out mutants were planted with randomization blocking. All mutants except for the *shm* mutants were sprayed with 0.125X glufosinate and scored 6 DAS. The *shm* mutants were sprayed with 0.5X glufosinate and scored 6 and 14 DAS. *Student's t-test* scores were calculated in R between controls and mutants for any statistically significant changes in tolerance between mutants and controls (R Core Team 2013).

3.2.2 Results

Two sets of 440 accessions were sprayed with different concentrations of glufosinate, 0.125X and 0.25X. The damage was scored on a scale of 1-5, 1 representing

no harm and 5 representing death. The natural variation of 440 accessions was broad depending on the rate of application and time after the glufosinate application (Figure 3.1). Using GWA, associations were calculated between the phenotypes and the SNPs using EMMAX and MLMM to find genetic factors that influence tolerance in *A. thaliana*.

The GWA of these ten phenotypes were calculated using the pipeline as described in chapter 2. As mentioned in the introduction, the interpretation of the GWA results can be difficult, and determining the true positives from the spurious positives is difficult. For most of the phenotypes, there were no clear peaks of significance, and there were multiple regions of the genome that showed significance depending on which SNP data set was used and which statistical method was used (Figures 3.2-3.5). The GWA results from EMMAX and MLMM did not give clear and obvious candidate genes that contributed to the natural variance of glufosinate tolerance (Table 3.1). Therefore, we looked at candidate genes based on the biological mechanism of glufosinate. We analyzed the significance of SNPs linked to genes involved in nitrogen use, and photorespiration including the *GS* genes and the *SHM* genes to determine if these genes contributed to glufosinate tolerance.

Using preliminary data (not available), we selected eight putative genes based on their biological function (Table 3.2). These genes were involved in cutin biosynthesis, nitrogen use, stress response, or sugar-induced gene expression. We hypothesized that the cutin biosynthesis could contribute to inhibiting glufosinate from entering the cell. We hypothesized that changes in glucosinolate and amino acid biosynthesis could ameliorate the effects of the inhibited GS enzymes. We hypothesized that inducing the stress response in plants could contribute to more glufosinate tolerance. Lastly, we

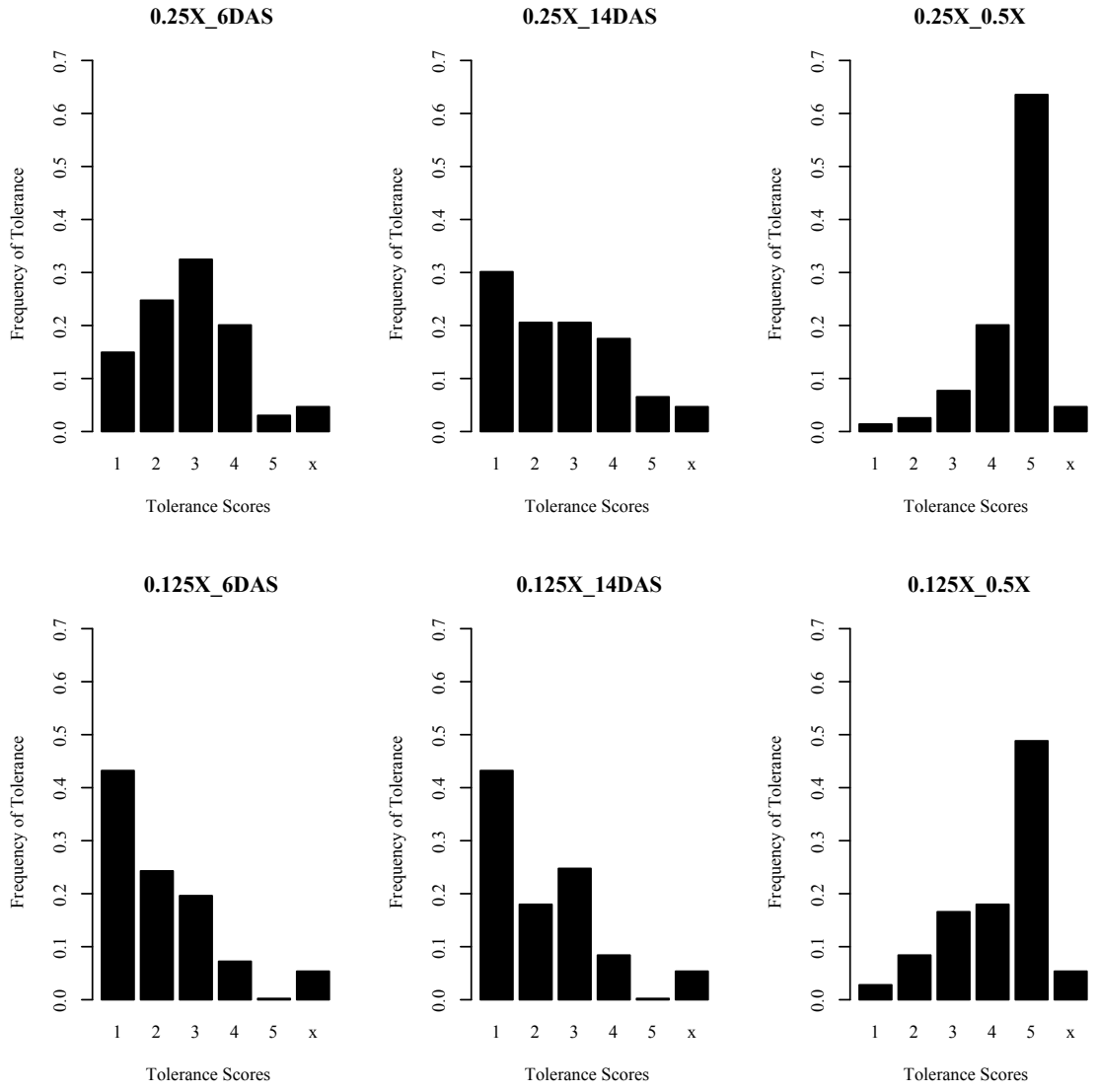

Figure 3.1 The natural variation of glufosinate tolerance in 440 accessions of *A. thaliana* is displayed using the frequencies of each score for each phenotype as shown in the bar graphs. The y-axis is the frequency of plants for each tolerance score. The x-axis explains the score given each plant: 1 = no glufosinate damage; 2= 1-2 leaves show damage; 3= 3-4 leaves show damage; 4=all leaves show damage and only meristematic tissue is alive; 5= the plant is dead.

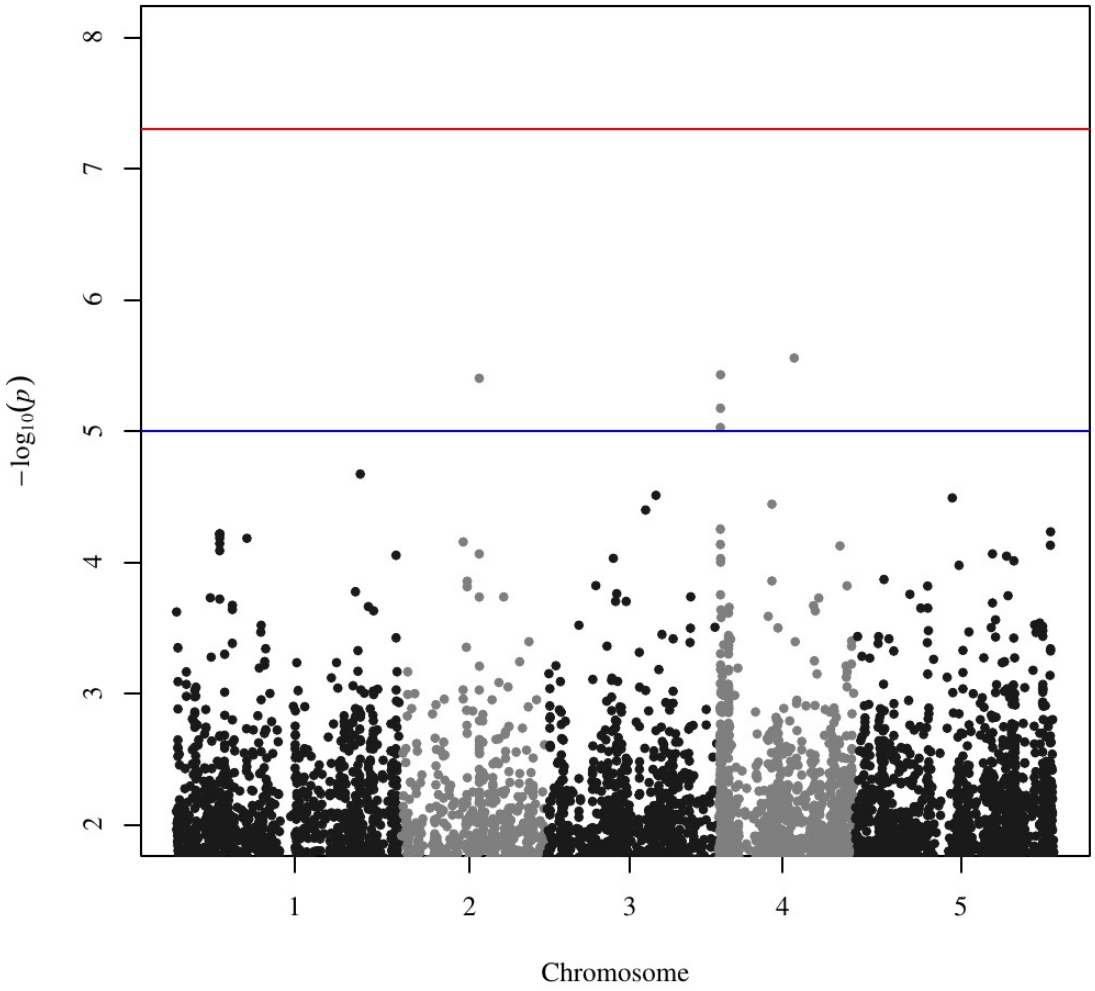

Figure 3.2. Manhattan plot displaying the *p*-values of 211K SNPs calculated using EMMAX for the phenotype Grand Mean. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The blue line indicates suggested significant cutoff ($\alpha \leq 1x10^{-5}$) and the red line indicates a genomewide significant cutoff ($\alpha \leq 5x10^{-8}$).

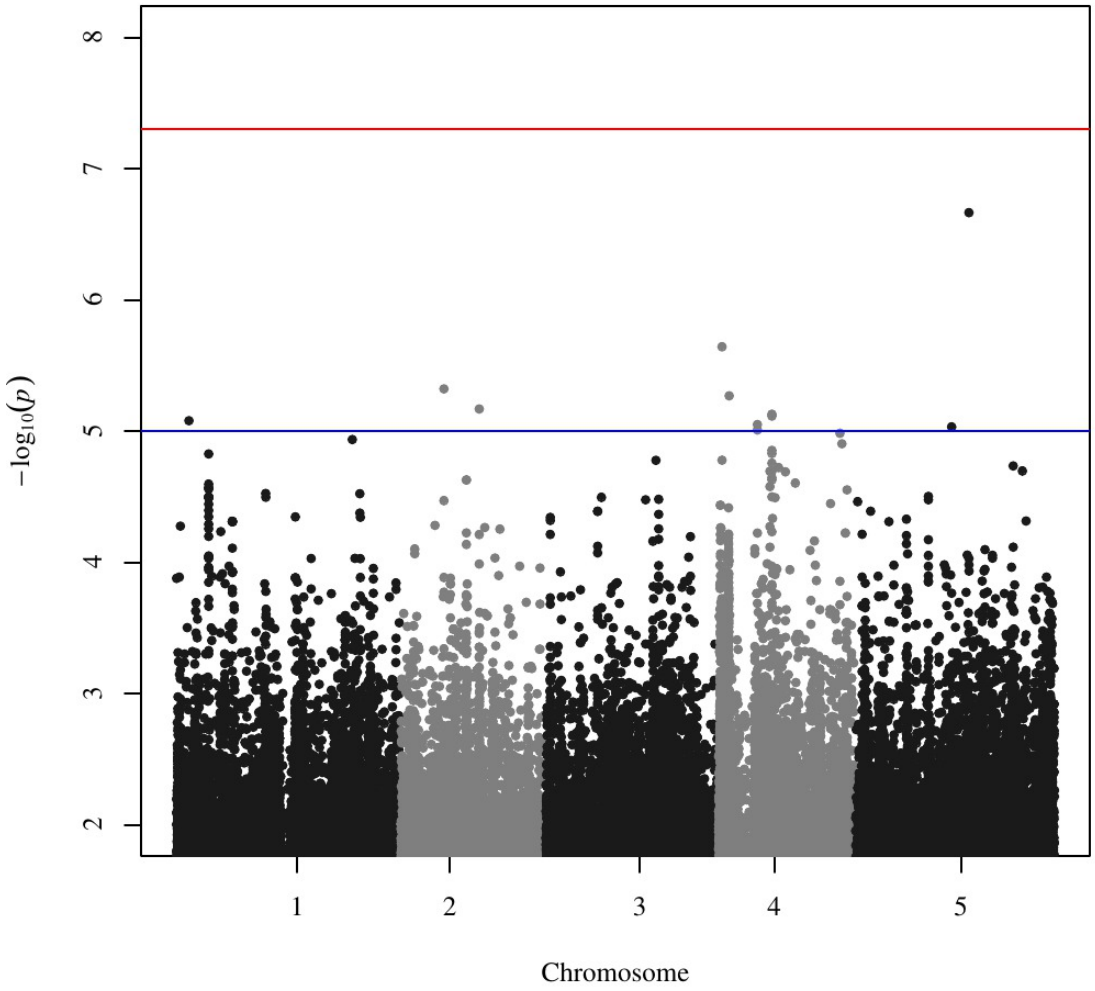

Figure 3.3. Manhattan plot displaying the *p*-values of 1.6M SNPs calculated using EMMAX for the phenotype Grand Mean. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The blue line indicates suggested significant cutoff ($\alpha \leq 1x10^{-5}$) and the red line indicates a genomewide significant cutoff ($\alpha \leq 5x10^{-8}$).

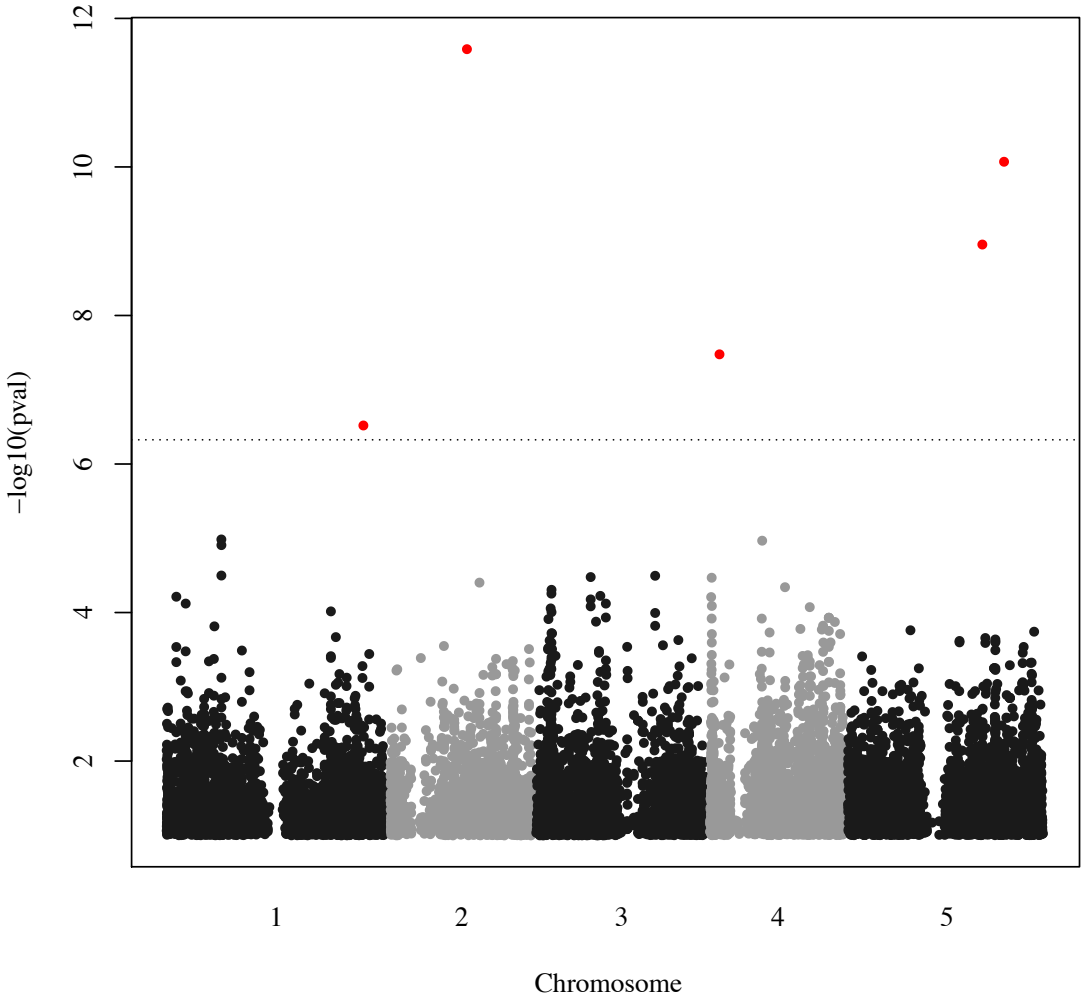

Figure 3.4. Manhattan plot displaying the *p*-values of 211K SNPs calculated using MLMM for the phenotype Grand Mean. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The red highlighted SNPs are the significant SNPs as determined by the model selection criterion BIC.

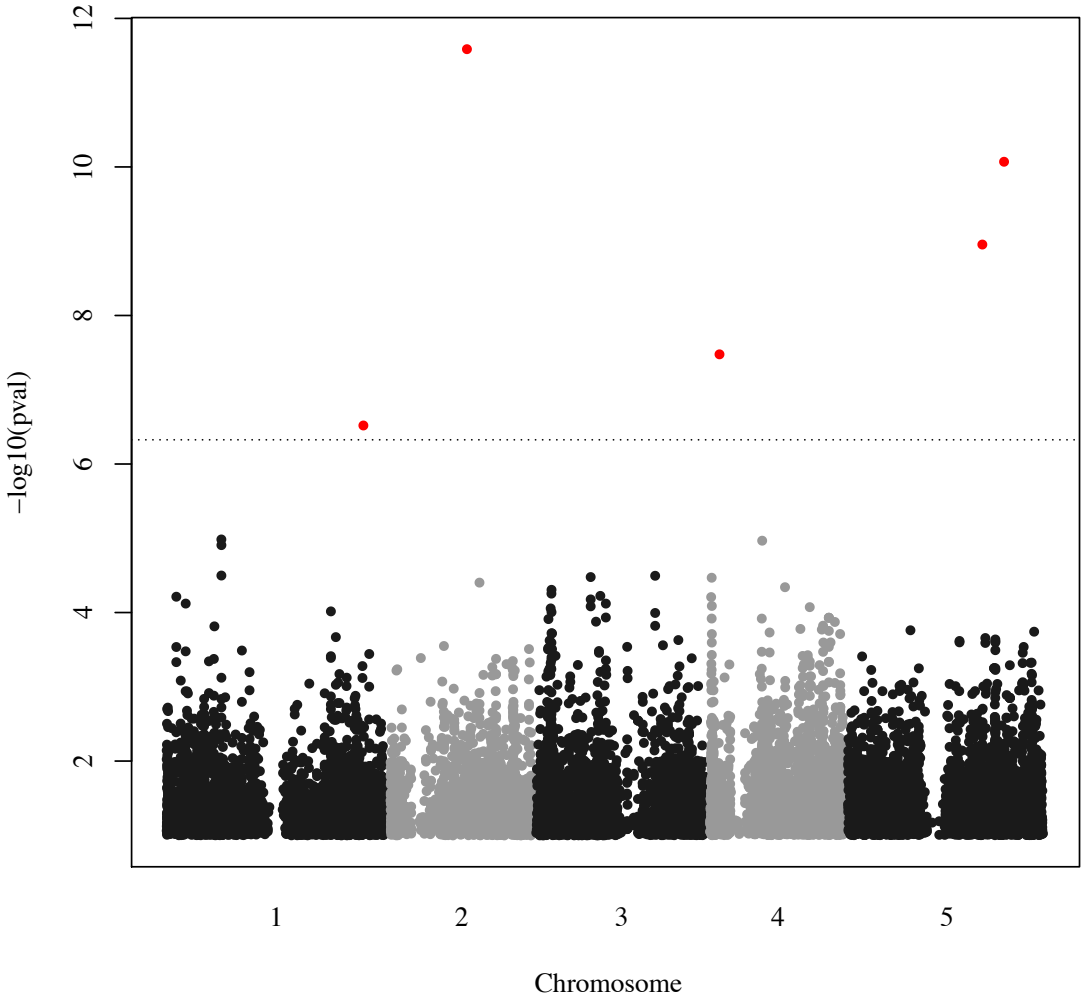

Figure 3.5. Manhattan plot displaying the *p*-values of 211K SNPs calculated using MLMM for the phenotype Grand Mean. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The red highlighted SNPs are the significant SNPs as determined by the model selection criterion Bonferroni.

Table 3.1. The number of significant SNPs from the two statistical models: EMMAX and MLMM for glufosinate tolerance. Significance for the EMMAX results was defined as ($\alpha \leq 1x10^{-5}$).

Phenotype	EMMAX 211K	EMMAX 1.6M	MLMM BIC	MLMM BONF
Grand Mean	5	11	5	5
Mean_6DAS	0	11	2	2
Mean_14DAS	5	15	4	9
Mean_0.5X	4	44	3	3
0.25X_0.5X	19	48	2	3
0.25X_6DAS	1	4	2	2
0.25X_14DAS	1	14	0	1
0.125X_0.5X	3	31	2	6
0.125X_6DAS	6	34	2	2
0.125X_14DAS	9	23	6	6

Table 3.2. The eight candidate genes for glufosinate tolerance selected by biological function. The eight genes are involved in amino acid metabolism, glucosinolate biosynthesis, stress response, or cutin biosynthesis. The Phenotype column describes which phenotype the significant SNPs close to or within the candidate genes were found using the 211K SNPs dataset, except for K23L20.6, which was found using the 1.6M SNPs dataset. SNPs. *This gene was found in the list of candidate genes using 1.6M

Gene	Gene Name	Mutant	Phenotype	Gene Function	References
AT1G17290	*ALANINE AMINOTRANSFERASE*	*alaat1-1*	Grand Mean, Mean_14DAS, 0.25X_14DAS, 0.125X_14DAS, 0.125X_6DAS	Amino acid metabolism	(Liepman and Olsen 2003, Good et al. 2007)
AT2G30470	*HIGH-LEVEL EXPRESSION OF SUGAR-INDUCIBLE GENE2*	*hsi2*	0.25X_6DAS, 0.125X_14DAS	Transcription repressor	(Tsukagoshi et al. 2005)
AT3G45300	*ISOVALERYL-COA-DEHYDROGENASE*	***ivd1-1***		Amino acid metabolism	(Däschner et al. 1999, 2001, Gu et al. 2010, Araújo et al. 2010)
AT4G00400	*GLYCEROL-3-PHOSPHATE SN-2-ACYLTRANSFERASE8*	*gpat8*	0.125X_14DAS, 0.125X_14DAS	Cutin biosynthesis	(Foy 1964, Bukovac and Petracek 1993, Li et al. 2007)
AT4G03050	*2-OXOGLUTARATE-DEPENDENT DIOXYGENASE*	*aop3*	0.125X_6DAS Mean_0.5X	Glucosinolate biosynthesis	(Kliebenstein et al. 2001, Grubb and Abel 2006, Chan et al. 2010, Sønderby et al. 2010)
AT4G03070	*2-OXOGLUTARATE-DEPENDENT DIOXYGENASE*	*aop1*	0.125X_6DAS Mean_0.5X	Glucosinolate biosynthesis	(Kliebenstein et al. 2001, Grubb and Abel 2006, Chan et al. 2010, Sønderby et al. 2010)
AT5G44720	*MOLYBDENUM COFACTOR SULFURASE*	*k23l20.6*	0.25X_0.5X*	Role in metabolic processes	(Gupta et al. 1991, Stallmeyer et al. 1999, Mendel 2002, Liu et al. 2009, Ide et al. 2011)
AT5G53120	*SPERMIDINE SYNTHASE3*	*spds3*	0.125X_6DAS	Stress Response	(Bagni and Tassoni 2001, Kasukabe et al. 2004, Yamaguchi et al. 2007, Broz et al. 2008, Zhao et al. 2010, Zhang et al. 2013)

hypothesized that changing sugar-inducible gene expression could ameliorate the effects of decreased Calvin Cycle output.

3.2.2.1 Determining the significance of candidate genes

To determine if the candidate genes contributed to glufosinate tolerance we tested if the SNPs within or in a 20kb window of the candidate genes would have significant associations. We used the EMMAX output from the two SNP datasets, 211K and 1.6M, to find the most significant SNP within the gene and within the 20 kb window of each gene. we calculated a new significant cutoff using the Bonferroni method ($\alpha \leq 0.05/\#$ SNPs) for each gene. This was less stringent than the cutoff used for the GWA since a smaller number of hypotheses was tested for each gene.

3.2.2.1.1 Glutamine synthetase

GS is the target enzyme for glufosinate. The six paralogs have different functions within the plant, GS2 being the main enzyme for photorespiration. The number of SNPs within the genes using the 211K SNPs dataset was very small, and none of the genes showed any significance using the Bonferroni cutoff (Table 3.3). Expanding the window to a 10kb upstream and downstream of the genes increased the number of SNPs analyzed, but none of the SNPs had a significant association between phenotype and SNP (Table 3.4).

Using the 1.6M SNPs dataset, the number of SNPs increased within each gene, and a significant SNP was found in *GS1-2* in the 0.25X_6DAS phenotype (Table 3.5), and upon expanding to the 20kb window, a SNP close to *GS1-4* was significant in the 0.25X_14DAS phenotype (Table 3.6). The lack of significant association between SNPs

Table 3.3 The top SNPs within the six paralogs of *GS* for each phenotype from the EMMAX model using 211K SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. None of the SNPs were found to be significant after the Bonferroni correction.

	GS1-1		***GS1-2***		***GS1-3***	
	SNP	P	SNP	P	SNP	P
Grand Mean	514935615	0.045[a]	124656612	0.512	36098503	0.274
Mean_14DAS	514933682	0.068	124656612	0.430	36098503	0.465
Mean_6DAS	514933682	0.093	124656612	0.987	36098503	0.249
Mean_0.5X	514935615	0.025[a]	124656612	0.566	36098732	0.616
0.25X_14DAS	514935358	0.049[a]	124656612	0.348	36098732	0.724
0.25X_6DAS	514933486	0.231	124656612	0.656	36098503	0.267
0.25X_0.5X	514935615	0.048[a]	124656612	0.077	36098503	0.028[a]
0.125X_14DAS	514933682	0.048[a]	124656612	0.680	36098503	0.314
0.125X_6DAS	514933682	0.060	124656612	0.894	36098732	0.321
0.125X_0.5X	514935336	0.017[a]	124656612	0.699	36098503	0.314
n	15		1		2	
Bonf	0.003		0.050		0.025	
	GS1-4		***GS1-5***		***GS2***	
	SNP	P	SNP	P	SNP	P
Grand Mean	55422801	0.271	117914148	0.119	513832505	0.311
Mean_14DAS	55422801	0.153	117914148	0.289	513833427	0.351
Mean_6DAS	55422801	0.308	117914148	0.105	513832505	0.115
Mean_0.5X	55423924	0.328	117913720	0.322	513831381	0.180
0.25X_14DAS	55422801	0.099	117913720	0.364	513832505	0.496
0.25X_6DAS	55422801	0.215	117914148	0.260	513832505	0.432
0.25X_0.5X	55423924	0.047[a]	117914148	0.147	513832505	0.155
0.125X_14DAS	55423065	0.020[a]	117914148	0.372	513833427	0.144
0.125X_6DAS	55423924	0.045[a]	117914148	0.115	513832505	0.081
0.125X_0.5X	55423340	0.533	117914148	0.461	513831381	0.505
n	5		4		4	
Bonf	0.010		0.013		0.013	

Table 3.4. The top SNPs within and 10kb up and downstream the six paralogs of *GS* for each phenotype from the EMMAX model using 211K SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. None of the SNPs were found to be significant after the Bonferroni correction.

	GS1-1		***GS1-2***		***GS1-3***	
	SNP	P	SNP	P	SNP	P
Grand Mean	514926587	0.020[a]	124666155	0.005[a]	36103781	0.054
Mean_14DAS	514936848	0.052	124655043	0.005[a]	36107246	0.027[a]
Mean_6DAS	514933682	0.093	124666155	0.003[a]	36107604	0.016[a]
Mean_0.5X	514926587	0.003[a]	124648147	0.037[a]	36107604	0.177
0.25X_14DAS	514926720	0.068	124666747	0.002[a]	36090938	0.098
0.25X_6DAS	514928814	0.056	124666155	0.014[a]	36107604	0.025[a]
0.25X_0.5X	514924618	0.013[a]	124667387	0.013[a]	36098503	0.028[a]
0.125X_14DAS	514933682	0.048[a]	124652362	0.044[a]	36109505	0.002[a]
0.125X_6DAS	514924212	0.032[a]	124666155	0.005[a]	36107604	0.139
0.125X_0.5X	514926587	0.009[a]	124654687	0.017[a]	36091661	0.026[a]
n	64		33		37	
Bonf	0.001		0.002		0.001	
	GS1-4		***GS1-5***		***GS2***	
	SNP	P	SNP	P	SNP	P
Grand Mean	55426858	0.088	117920697	0.013[a]	513822767	0.006[a]
Mean_14DAS	55427408	0.073	117920697	0.034[a]	513823553	0.007[a]
Mean_6DAS	55426858	0.082	117925713	0.013[a]	513822087	0.010[a]
Mean_0.5X	55419248	0.019[a]	117925326	0.013[a]	513822767	0.021[a]
0.25X_14DAS	55422801	0.099	117904307	0.076	513823261	0.007[a]
0.25X_6DAS	55426858	0.028[a]	117925713	0.029[a]	513822087	0.005[a]
0.25X_0.5X	55419248	0.032[a]	117925326	0.011[a]	513822767	0.026[a]
0.125X_14DAS	55416621	0.007[a]	117920697	0.018[a]	513823553	0.003[a]
0.125X_6DAS	55416621	0.023[a]	117925326	0.058	513823553	0.013[a]
0.125X_0.5X	55433648	0.080	117922960	0.078	513822767	0.062[a]
n	24		39		35	
Bonf	0.002		0.001		0.001	

Table 3.5. The top SNPs within the six paralogs of *GS* for each phenotype from the EMMAX model using 1.6M SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. None of the SNPs were found to be significant after the Bonferroni correction.

	GS1-1		***GS1-2***		***GS1-3***	
	SNP	P	SNP	P	SNP	P
Grand Mean	514933682	0.029[a]	124655978	0.013[a]	36098866	0.148
Mean_14DAS	514933682	0.023[a]	124656812	0.008[a]	36099551	0.326
Mean_6DAS	514933682	0.053	124655978	0.059	36098592	0.119
Mean_0.5X	514933682	0.027[a]	124655978	0.106	36098072	0.320
0.25X_14DAS	514933799	0.013[a]	124656812	0.002[b]	36098866	0.191
0.25X_6DAS	514933996	0.075	124655978	0.074	36098866	0.216
0.25X_0.5X	514933799	0.047[a]	124656838	0.098	36098503	0.043[a]
0.125X_14DAS	514934478	0.013[a]	124657177	0.105	36098059	0.400
0.125X_6DAS	514934912	0.021[a]	124655493	0.041[a]	36098592	0.016[a]
0.125X_0.5X	514935401	0.034[a]	124655978	0.042[a]	36098503	0.338
n	77		18		14	
Bonf	0.001		0.003		0.004	
	GS1-4		***GS1-5***		***GS2***	
	SNP	P	SNP	P	SNP	P
Grand Mean	55424115	0.034[a]	117914148	0.121	513831196	0.034[a]
Mean_14DAS	55424115	0.010[a]	117914148	0.396	513830606	0.008[a]
Mean_6DAS	55424100	0.150	117914148	0.104	513832505	0.045[a]
Mean_0.5X	55424115	0.031[a]	117914048	0.523	513830640	0.021[a]
0.25X_14DAS	55422332	0.005[a]	117914048	0.286	513830606	0.012[a]
0.25X_6DAS	55421793	0.067	117914048	0.312	513831196	0.066
0.25X_0.5X	55422332	0.058	117914148	0.158	513830640	0.038[a]
0.125X_14DAS	55423065	0.011[a]	117914411	0.170	513831196	0.103
0.125X_6DAS	55423602	0.076	117914148	0.069	513832505	0.033[a]
0.125X_0.5X	55424115	0.007[a]	117914048	0.331	513830640	0.231[a]
n	12		9		39	
Bonf	0.004		0.006		0.001	

Table 3.6. The top SNPs within and 10kb up and downstream the six paralogs of *GS* for each phenotype from the EMMAX model using 1.6M SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. None of the SNPs were found to be significant after the Bonferroni correction.

	GS1-1		***GS1-2***		***GS1-3***	
	SNP	P	SNP	P	SNP	P
Grand Mean	514926785	0.003[a]	124665726	0.007[a]	36092707	0.017[a]
Mean_14DAS	514933682	0.023[a]	124655043	0.002[a]	36107809	0.005[a]
Mean_6DAS	514939290	0.003[a]	124666778	0.004[a]	36105476	0.023[a]
Mean_0.5X	514926785	0.000[a]	124645638	0.005[a]	36101196	0.027[a]
0.25X_14DAS	514933799	0.013[a]	124666747	0.001[a]	36107809	0.005[a]
0.25X_6DAS	514938628	0.006[a]	124664104	0.001[a]	36107604	0.046[a]
0.25X_0.5X	514933313	0.001[a]	124659412	0.001[a]	36105851	0.008[a]
0.125X_14DAS	514933212	0.010[a]	124662013	0.009[a]	36108157	0.002[a]
0.125X_6DAS	514924225	0.001[a]	124666807	0.006[a]	36105476	0.009[a]
0.125X_0.5X	514926785	0.001[a]	124654715	0.018[a]	36101196	0.021[a]
n	424		343		226	
Bonf	0.000		0.000		0.000	
	GS1-4		***GS1-5***		***GS2***	
	SNP	P	SNP	P	SNP	P
Grand Mean	55424115	0.034[a]	117920697	0.011[a]	513823553	0.001[a]
Mean_14DAS	55430065	0 .007[a]	117920697	0.039[a]	513823553	0.001[a]
Mean_6DAS	55430065	0.024[a]	117923408	0.019[a]	513823553	0.005[a]
Mean_0.5X	55428056	0.029[a]	117923953	0.006[a]	513842280	0.001[a]
0.25X_14DAS	55428628	0.000[b]	117921395	0.054	513834271	0.003[a]
0.25X_6DAS	55430139	0.020[a]	117907895	0.021[a]	513822087	0.004[a]
0.25X_0.5X	55417523	0.008[a]	117923938	0.001[a]	513842280	0.001[a]
0.125X_14DAS	55416659	0.004[a]	117904028	0.006[a]	513823553	0.001[a]
0.125X_6DAS	55431533	0.026[a]	117903941	0.015[a]	513820866	0.003[a]
0.125X_0.5X	55424115	0.007[a]	117903941	0.029[a]	513827731	0.014[a]
n	130		108		335	
Bonf	0.000		0.000		0.000	

linked to the *GS* genes and the phenotypes suggested that variation within GS enzymes was not contributing to glufosinate tolerance in *A. thaliana.*

3.2.2.1.2 Serine hydroxymethyltransferase

The SHM family consists of seven enzymes that convert glycine to serine (Somerville and Ogren 1981). SHM1 is the enzyme that participates in photorespiration, but it was recently found that SHM3 localizes in plastids and potentially could have a role in photorespiration also (McClung et al. 2000, Voll et al. 2006, Zhang et al. 2010). In the 211K SNPs dataset, none of the SNPs were within the *SHM1* gene. In the 1.6M SNPs dataset the SNPs within *SHM1* did not show any significance (Table 3.9). *SHM3* showed significance after a Bonferroni correction for the phenotypes Mean_6DAS, 0.125X_14DAS, and 0.125X_6DAS for the 211K SNPs dataset and for the phenotypes Mean_6DAS and 0.125X_6DAS for the 1.6M SNPs dataset (Tables 3.7 & 3.9). The other genes did not show any significance after the Bonferroni correction for both SNP datasets.

Upon expanding our search to 10kb upstream and downstream the genes, SNPs associated with *SHM3* and *SHM4* showed significance in differing phenotypes in the 211K SNPs dataset, and the only common phenotype was Mean_6DAS (Table 3.8). The SNPs that were significant after a Bonferroni correction were found within *SHM3*. The SNPs for *SHM4* were found outside of the gene (Table 3.8).

Looking at the 1.6M SNPs dataset, *SHM4* and *SHM6* had a significant SNP associated with each gene. A SNP downstream of *SHM4* was significant in the Mean_6DAS phenotype. A SNP downstream of *SHM6* showed significance in the 0.125X_14DAS phenotype (Table 3.10). Otherwise, variation within the other genes did

Table 3.7. The top SNPs within the seven paralogs of *SHM* for each phenotype from the EMMAX model using 211K SNPs. There were no SNPs of SHM1. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. The subscript 'b' indicates SNPs that are significant with the Bonferroni correction.

	SHM2		***SHM3***		***SHM4***	
	SNP	P	SNP	P	SNP	P
Grand Mean	59421853	0.349	415689569	0.005[a]	48048160	0.078
Mean_14DAS	59420529	0.124	415692006	0.023[a]	48049683	0.086
Mean_6DAS	59420529	0.110	415692006	0.000[b]	48048629	0.119
Mean_0.5X	59421853	0.260	415692006	0.080	48049683	0.008[a]
0.25X_14DAS	59420529	0.329	415689569	0.273	48048614	0.016[a]
0.25X_6DAS	59421853	0.184	415692006	0.012[a]	48048614	0.109
0.25X_0.5X	59421853	0.045[a]	415692006	0.129	48048614	0.019[a]
0.125X_14DAS	59420529	0.168	415692006	0.002[b]	48048481	0.113
0.125X_6DAS	59419295	0.192	415692006	0.000[b]	48048160	0.058
0.125X_0.5X	59421880	0.529	415692006	0.232	48049683	0.011[a]
n	11		10		8	
Bonf	0.005		0.005		0.006	
	SHM5		***SHM6***		***SHM7***	
	SNP	P	SNP	P	SNP	P
Grand Mean	48032601	0.086	17754490	0.109	113696495	0.624
Mean_14DAS	48032601	0.263	17756594	0.065	113696739	0.506
Mean_6DAS	48032783	0.055	17754490	0.123	113697302	0.355
Mean_0.5X	48032601	0.135	17756296	0.242	113698247	0.177
0.25X_14DAS	48032783	0.142	17755321	0.096	113696739	0.566
0.25X_6DAS	48032783	0.024[a]	17756594	0.027[a]	113697302	0.477
0.25X_0.5X	48033275	0.227	17754567	0.168	113696495	0.535
0.125X_14DAS	48032601	0.055	17756594	0.029[a]	113697302	0.651
0.125X_6DAS	48032601	0.122	17754490	0.081	113697302	0.516
0.125X_0.5X	48032601	0.021[a]	17754490	0.097	113698247	0.172
n	8		8		4	
Bonf	0.006		0.006		0.013	

Table 3.8. The top SNPs within and 10kb up and downstream the seven paralogs of *SHM* for each phenotype from the EMMAX model using 211K SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. The subscript 'b' indicates SNPs that are significant with the Bonferroni correction.

	SHM1		***SHM2***		***SHM3***	
	SNP	P	SNP	P	SNP	P
Grand Mean	417823919	0.027[a]	59411427	0.105	415702807	0.004[a]
Mean_14DAS	417837235	0.104	59424671	0.090	415692006	0.023[a]
Mean_6DAS	417823919	0.006[a]	59411427	0.019[a]	415692006	0.000[b]
Mean_0.5X	417835086	0.140	59408646	0.038[a]	415679645	0.041[a]
0.25X_14DAS	417823125	0.079	59424671	0.019[a]	415681026	0.156
0.25X_6DAS	417823919	0.003[a]	59416728	0.016[a]	415692006	0.012[a]
0.25X_0.5X	417823125	0.024[a]	59421853	0.045[a]	415681669	0.038[a]
0.125X_14DAS	417822597	0.040[a]	59423051	0.094	415692006	0.002[a]
0.125X_6DAS	417837235	0.011[a]	59425504	0.047[a]	415692006	0.000[b]
0.125X_0.5X	417830131	0.174	59408646	0.041[a]	415679645	0.042[a]
n	19		77		54	
Bonf	0.003		0.001		0.001	
	SHM4		***SHM5***		***SHM6***	
	SNP	P	SNP	P	SNP	P
Grand Mean	48052654	0.000[b]	48034690	0.004[a]	17766920	0.015[a]
Mean_14DAS	48052654	0.000[b]	48034690	0.008[a]	17752449	0.047[a]
Mean_6DAS	48052654	0.000[b]	48031702	0.019[a]	17766920	0.039[a]
Mean_0.5X	48054930	0.002[a]	48034690	0.008[a]	17748122	0.012[a]
0.25X_14DAS	48052654	0.000[b]	48031702	0.017[a]	17752363	0.032[a]
0.25X_6DAS	48052654	0.000[b]	48031702	0.008[a]	17756594	0.027[a]
0.25X_0.5X	48048614	0.019[a]	48026327	0.009[a]	17748122	0.003[a]
0.125X_14DAS	48052654	0.002[a]	48022061	0.018[a]	17756594	0.029[a]
0.125X_6DAS	48053750	0.006[a]	48037404	0.007[a]	17766920	0.006[a]
0.125X_0.5X	48049683	0.011[a]	48034354	0.010[a]	17766920	0.005[a]
n	67		66		65	
Bonf	0.001		0.001		0.001	

Table 3.8 Continued

	SHM7	
	SNP	P
Grand Mean	113698746	0.026[a]
Mean_14DAS	113698746	0.031[a]
Mean_6DAS	113698746	0.047[a]
Mean_0.5X	113693311	0.053
0.25X_14DAS	113698746	0.067
0.25X_6DAS	113693388	0.049[a]
0.25X_0.5X	113693311	0.038[a]
0.125X_14DAS	113691854	0.029[a]
0.125X_6DAS	113699431	0.013[a]
0.125X_0.5X	113698746	0.157
n	42	
Bonf	0.001	

Table 3.9. The top SNPs within the seven paralogs of *SHM* for each phenotype from the EMMAX model using 1.6M SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. The subscript 'b' indicates SNPs that are significant with the Bonferroni correction.

	SHM1		***SHM2***		***SHM3***	
	SNP	P	SNP	P	SNP	P
Grand Mean	417832670	0.028[a]	59422156	0.061	415689569	0.025[a]
Mean_14DAS	417832758	0.051	59418402	0.034[a]	415689569	0.071
Mean_6DAS	417833773	0.011[a]	59422156	0.045[a]	415692006	0.001[b]
Mean_0.5X	417834603	0.020[a]	59421509	0.078	415691522	0.096
0.25X_14DAS	417832758	0.023[a]	59420456	0.072	415689580	0.241
0.25X_6DAS	417833773	0.001[a]	59420494	0.129	415690644	0.022[a]
0.25X_0.5X	417833773	0.010[a]	59421509	0.007[a]	415691311	0.033[a]
0.125X_14DAS	417832670	0.133	59421847	0.012[a]	415692006	0.005[a]
0.125X_6DAS	417832670	0.083	59422156	0.046[a]	415692006	0.002[b]
0.125X_0.5X	417834603	0.086	59420456	0.254	415689870	0.236
n	74		128		28	
Bonf	0.001		0.000		0.002	
	SHM4		***SHM5***		***SHM6***	
	SNP	P	SNP	P	SNP	P
Grand Mean	48048509	0.017[a]	48033463	0.047[a]	17757091	0.022[a]
Mean_14DAS	48048509	0.041[a]	48033463	0.026[a]	17757091	0.041[a]
Mean_6DAS	48050055	0.034[a]	48032783	0.010[a]	17757091	0.005[a]
Mean_0.5X	48049683	0.006[a]	48032601	0.090	17757091	0.049[a]
0.25X_14DAS	48050055	0.036[a]	48033463	0.021[a]	17754826	0.218
0.25X_6DAS	48048614	0.014[a]	48032783	0.008[a]	17757091	0.009[a]
0.25X_0.5X	48048509	0.002[a]	48033545	0.016[a]	17757091	0.077
0.125X_14DAS	48050030	0.036[a]	48032601	0.114	17757091	0.004[a]
0.125X_6DAS	48050030	0.009[a]	48033035	0.049[a]	17757091	0.025[a]
0.125X_0.5X	48049683	0.022[a]	48032601	0.026[a]	17754451	0.099
n	54		29		16	
Bonf	0.001		0.002		0.003	

Table 3.9. Continued

	SHM7	
	SNP	P
Grand Mean	113696719	0.129
Mean_14DAS	113696511	0.075
Mean_6DAS	113696531	0.025[a]
Mean_0.5X	113696962	0.051
0.25X_14DAS	113696511	0.154
0.25X_6DAS	113696531	0.022[a]
0.25X_0.5X	113696731	0.089
0.125X_14DAS	113696719	0.083
0.125X_6DAS	113696719	0.120
0.125X_0.5X	113696122	0.049[a]
n	33	
Bonf	0.002	

Table 3.10. The top SNPs within and 10kb up and downstream the six paralogs of *SHM* for each phenotype from the EMMAX model using 1.6M SNPs. The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. The subscript 'b' indicates SNPs that are significant with the Bonferroni correction.

	SHM1		***SHM2***		***SHM3***	
	SNP	P	SNP	P	SNP	P
Grand Mean	417838373	0.017[a]	59409877	0.012[a]	415702807	0.006[a]
Mean_14DAS	417837703	0.028[a]	59423852	0.016[a]	415702807	0.019[a]
Mean_6DAS	417823919	0.008[a]	59422681	0.006[a]	415692006	0.001[a]
Mean_0.5X	417837674	0.003[a]	59425435	0.026[a]	415680408	0.001[a]
0.25X_14DAS	417837703	0.010[a]	59424671	0.019[a]	415698707	0.005[a]
0.25X_6DAS	417833773	0.001[a]	59410970	0.013[a]	415680007	0.002[a]
0.25X_0.5X	417830575	0.005[a]	59421509	0.007[a]	415696518	0.000[a]
0.125X_14DAS	417824664	0.029[a]	59423852	0.006[a]	415687241	0.003[a]
0.125X_6DAS	417837235	0.006[a]	59422681	0.015[a]	415695873	0.001[a]
0.125X_0.5X	417842527	0.033[a]	59409877	0.012[a]	415680408	0.002[a]
n	288		615		274	
Bonf	0.000		0.000		0.000	
	SHM4		***SHM5***		***SHM6***	
	SNP	P	SNP	P	SNP	P
Grand Mean	48053750	0.001[a]	48036829	0.018[a]	17761169	0.012[a]
Mean_14DAS	48053750	0.001[a]	48028575	0 .018[a]	17759622	0.007[a]
Mean_6DAS	48053750	0.000[b]	48031367	0.008[a]	17757091	0.005[a]
Mean_0.5X	48052835	0.000[a]	48022255	0.001[a]	17744943	0.005[a]
0.25X_14DAS	48052391	0.002[a]	48031821	0.001[a]	17752449	0.006[a]
0.25X_6DAS	48052654	0.001[a]	48032783	0.008[a]	17746951	0.003[a]
0.25X_0.5X	48052835	0.001[a]	48036219	0.008[a]	17749852	0.010[a]
0.125X_14DAS	48051333	0.006[a]	48036217	0.006[a]	17759622	0.000[b]
0.125X_6DAS	48053750	0.001[a]	48034749	0.002[a]	17751442	0.003[a]
0.125X_0.5X	48052835	0.003[a]	48022368	0.001[a]	17744943	0.003[a]
n	488		473		347	
Bonf	0.000		0.000		0.000	

Table 3.10. Contintued.

	SHM7	
	SNP	P
Grand Mean	113687254	0.013[a]
Mean_14DAS	113690664	0 .011[a]
Mean_6DAS	113695244	0.003[a]
Mean_0.5X	113688424	0.005[a]
0.25X_14DAS	113704145	0.011[a]
0.25X_6DAS	113686418	0.003[a]
0.25X_0.5X	113708341	0.060
0.125X_14DAS	113695185	0.003[a]
0.125X_6DAS	113689417	0.004[a]
0.125X_0.5X	113687765	0.008[a]
n	438	
Bonf	0.000	

not show any significance. Even though a few SNPs did show some significance the overall evidence suggested that the *SHM* genes did not contribute to glufosinate tolerance.

3.2.2.1.3 Photorespiration

I looked at the SNPs within 37 photorespiration genes. Ten of the genes showed significance in at least one phenotype after the Bonferroni correction, either within the gene or within a 20kb window of the gene in the 211K SNPs dataset (Tables 3.11 & 3.12). A SNP within the gene *FERREDOXIN-DEPENDENT GLUTAMATE SYNTHASE* (*GLS*), which encodes the enzyme ferredoxin-depending glutamine:2-oxoglutarate amidotransferase (FD-GOGAT) was significant for the overall mean score, Grand Mean (Table 3.11). FD-GOGAT is the second enzyme in the nitrogen cycle of converting ammonia and glutamate into glutamine. FD-GOGAT binds glutamine and α-ketoglutarate to make glutamine (Coschigano et al. 1998, Suzuki and Knaff 2005). This could indicate that though the effect of the variation within *GLS* was too small to capture using GWA in any one phenotype, overall the gene did have an affect on the amount of tissue damage caused by glufosinate.

In the 1.6M SNPs dataset, six genes showed significance in at least one phenotype within or surrounding the photorespiration genes. None of the SNPs within or linked to *GLS* is significant after the Bonferroni correction (Tables 3.13 & 3.14). Once again, the genetic variation of photorespiration genes did not appear to play a role in glufosinate tolerance.

Table 3.11. The top SNPs within genes involved in photorespiration for each phenotype from the EMMAX model using 211K SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. The subscript 'b' indicates SNPs that are significant with the Bonferroni correction.

	AT1G11860		**AT1G14450**		**AT1G23310**	
	SNP	P	SNP	P	SNP	P
Grand Mean	14001906	0.101	14947471	0.128	18271200	0.677
Mean_14DAS	14001705	0.305	14947128	0.037[a]	18271200	0.735
Mean_6DAS	14001705	0.570	14947471	0.097	18269549	0.399
Mean_0.5X	14001906	0.015[a]	14946664	0.245	18269549	0.354
0.25X_14DAS	14001906	0.117	14946664	0.047[a]	18269549	0.227
0.25X_6DAS	14001906	0.040[a]	14946664	0.132	18269549	0.524
0.25X_0.5X	14001906	0.023[a]	14946664	0.350	18271200	0.904
0.125X_14DAS	14002668	0.053	14947128	0.035[a]	18269549	0.075
0.125X_6DAS	14002963	0.181	14947471	0.083	18271200	0.434
0.125X_0.5X	14003257	0.033[a]	14947471	0.574	18269549	0.166
n	7		3		2	
Bonf	0.007		0.017		0.025	
	AT1G32470		**AT1G48030**		**AT1G67350**	
	SNP	P	SNP	P	SNP	P
Grand Mean	111740236	0.087	117717717	0.052	125236635	0.088
Mean_14DAS	111740236	0.072	117717717	0.199	125236635	0.457
Mean_6DAS	111739738	0.145	117717630	0.098	125236360	0.352
Mean_0.5X	111740236	0.114	117718858	0.040[a]	125236635	0.009[b]
0.25X_14DAS	111740236	0.193	117717717	0.063	125236635	0.253
0.25X_6DAS	111739738	0.411	117717717	0.162	125236360	0.411
0.25X_0.5X	111740019	0.457	117717717	0.035[a]	125236635	0.010 [b]
0.125X_14DAS	111740236	0.053	117717630	0.074	125235748	0.383
0.125X_6DAS	111739738	0.220	117717630	0.010[b]	125236635	0.426
0.125X_0.5X	111740236	0.061	117717630	0.113	125236635	0.074
n	3		4		5	
Bonf	0.017		0.013		0.010	

Table 3.11

	AT1G68010		**AT1G70580**		**AT1G80380**	
	SNP	P	SNP	P	SNP	P
Grand Mean	125494350	0.118	126615577	0.573	130217346	0.198
Mean_14DAS	125494350	0.017[a]	126615577	0.288	130217346	0.054
Mean_6DAS	125494350	0.425	126612910	0.726	130217346	0.771
Mean_0.5X	125495137	0.136	126615577	0.685	130217346	0.168
0.25X_14DAS	125494350	0.060	126612910	0.474	130217346	0.086
0.25X_6DAS	125495987	0.515	126615577	0.588	130217346	0.605
0.25X_0.5X	125494350	0.168	126612910	0.579	130219655	0.258
0.125X_14DAS	125494350	0.075	126615577	0.318	130217346	0.096
0.125X_6DAS	125494350	0.665	126615577	0.481	130218486	0.416
0.125X_0.5X	125495137	0.185	126615577	0.791	130217346	0.121
n	3		2		3	
Bonf	0.017		0.025		0.017	
	AT2G02050		**AT2G04540**		**AT2G13360**	
	SNP	P	SNP	P	SNP	P
Grand Mean	2490568	0.984	21581526	0.028[a]	25539651	0.199
Mean_14DAS	2490568	0.955	21584422	0.040[a]	25540646	0.209
Mean_6DAS	2490568	0.567	21582157	0.242	25540226	0.039
Mean_0.5X	2490568	0.482	21581526	0.046[a]	25540646	0.105
0.25X_14DAS	2490568	0.941	21584422	0.016[a]	25540646	0.501
0.25X_6DAS	2490568	0.878	21582157	0.161	25539651	0.362
0.25X_0.5X	2490568	0.278	21582399	0.232	25540646	0.355
0.125X_14DAS	2490568	0.941	21581526	0.030[a]	25540646	0.170
0.125X_6DAS	2490568	0.544	21581526	0.213	25540226	0.009[a]
0.125X_0.5X	2490568	0.944	21581526	0.009[a]	25540646	0.119
n	1		19		6	
Bonf	0.050		0.003		0.008	

Table 3.11 Continued

	AT2G26080		AT2G27730		AT2G33220	
	SNP	P	SNP	P	SNP	P
Grand Mean	211110158	0.044[a]	211821683	0.066	214080129	0.154
Mean_14DAS	211110158	0.131	211821683	0.040[a]	214079135	0.320
Mean_6DAS	211109872	0.176	211821683	0.689	214079135	0.556
Mean_0.5X	211110158	0.011[a]	211821683	0.040[a]	214080129	0.022[b]
0.25X_14DAS	211110158	0.088	211821683	0.063	214079135	0.453
0.25X_6DAS	211110158	0.035[a]	211821683	0.508	214080129	0.931
0.25X_0.5X	211109872	0.007[a]	211821683	0.150	214080129	0.013[b]
0.125X_14DAS	211112939	0.188	211821683	0.123	214079135	0.236
0.125X_6DAS	211110738	0.074	211821683	0.624	214079135	0.145
0.125X_0.5X	211110158	0.102	211821683	0.056	214080129	0.142
n	9		3		2	
Bonf	0.006		0.017		0.025	
	AT2G35120		**AT2G41220**		**AT3G14415**	
	SNP	P	SNP	P	SNP	P
Grand Mean	214805846	0.072	217186417	0.036[a]	34820182	0.278
Mean_14DAS	214805846	0.201	217186417	0.093	34818332	0.317
Mean_6DAS	214805846	0.080	217186417	0.032[a]	34818332	0.246
Mean_0.5X	214805846	0.097	217186417	0.025[a]	34819661	0.127
0.25X_14DAS	214806445	0.234	217186417	0.208	34819230	0.281
0.25X_6DAS	214805846	0.220	217181989	0.317	34818332	0.243
0.25X_0.5X	214807148	0.099	217186417	0.203	34819102	0.029
0.125X_14DAS	214805846	0.161	217177677	0.151	34818332	0.134
0.125X_6DAS	214805846	0.064	217186417	0.014[a]	34819230	0.306
0.125X_0.5X	214805846	0.028[a]	217186417	0.020[a]	34820182	0.487
n	4		4		5	
Bonf	0.013		0.013		0.010	

Table 3.11. Continued

	AT3G14420		AT3G17240		AT3G18410	
	SNP	P	SNP	P	SNP	P
Grand Mean	34823150	0.602	35892091	0.065	36324384	0.356
Mean_14DAS	34823150	0.968	35891424	0.101	36324384	0.309
Mean_6DAS	34823150	0.872	35892168	0.118	36324050	0.293
Mean_0.5X	34823150	0.538	35892091	0.087	36323815	0.050
0.25X_14DAS	34823150	0.349	35891424	0.026[a]	36324384	0.311
0.25X_6DAS	34823150	0.814	35892091	0.176	36324384	0.593
0.25X_0.5X	34823150	0.578	35892091	0.302	36323815	0.103
0.125X_14DAS	34823150	0.231	35891340	0.043[a]	36324384	0.532
0.125X_6DAS	34823150	0.946	35892168	0.046[a]	36324050	0.077
0.125X_0.5X	34823150	0.113	35891424	0.102	36323815	0.170
n	1		8		3	
Bonf	0.050		0.006		0.017	
	AT3G54110		**AT4G16450**		**AT4G33010**	
	SNP	P	SNP	P	SNP	P
Grand Mean	320039917	0.192	49280248	0.007[b]	415930433	0.127
Mean_14DAS	320039917	0.249	49280248	0.073	415930433	0.099
Mean_6DAS	320039392	0.187	49280248	0.010[a]	415930433	0.021[a]
Mean_0.5X	320039828	0.175	49280248	0.069	415930433	0.003[a]
0.25X_14DAS	320039917	0.473	49280248	0.306	415929822	0.330
0.25X_6DAS	320040118	0.492	49280248	0.127	415930433	0.032[a]
0.25X_0.5X	320039828	0.138	49280248	0.198	415929822	0.002[b]
0.125X_14DAS	320039917	0.105	49280248	0.013[a]	415930433	0.034[a]
0.125X_6DAS	320039392	0.120	49280248	0.007[b]	415930433	0.071
0.125X_0.5X	320041003	0.070	49280248	0.047[a]	415927833	0.074
n	7		4		8	
Bonf	0.007		0.013		0.006	

Table 3.11. Continued

	AT4G35090		AT5G04140		AT5G06580	
	SNP	P	SNP	P	SNP	P
Grand Mean	416703115	0.054	51132202	0.004[b]	52011427	0.142
Mean_14DAS	416700480	0.022[a]	51132202	0.007[a]	52011427	0.096
Mean_6DAS	416700991	0.071	51132202	0.017[a]	52011427	0.185
Mean_0.5X	416703032	0.005[a]	51132202	0.041[a]	52014746	0.056
0.25X_14DAS	416700991	0.098	51132202	0.005[a]	52014249	0.258
0.25X_6DAS	416700991	0.040[a]	51132202	0.119	52011427	0.179
0.25X_0.5X	416703032	0.070	51132202	0.047[a]	52014746	0.121
0.125X_14DAS	416700480	0.033[a]	51132202	0.088	52011427	0.101
0.125X_6DAS	416702929	0.023[a]	51132202	0.017[a]	52011427	0.434
0.125X_0.5X	416703032	0.005[a]	51132202	0.141	52014249	0.045[a]
n	11		10		3	
Bonf	0.005		0.005		0.017	
	AT5G12860		**AT5G35630**		**AT5G47760**	
	SNP	P	SNP	P	SNP	P
Grand Mean	54060053	0.099	513832505	0.311	519342933	0.061
Mean_14DAS	54061791	0.093	513833427	0.351	519342933	0.160
Mean_6DAS	54060806	0.205	513832505	0.115	519342933	0.080
Mean_0.5X	54060053	0.081	513831381	0.180	519342933	0.014[b]
0.25X_14DAS	54061791	0.135	513832505	0.496	519342933	0.196
0.25X_6DAS	54060806	0.259	513832505	0.432	519342933	0.071
0.25X_0.5X	54060053	0.157	513832505	0.155	519342933	0.023[b]
0.125X_14DAS	54060053	0.112	513833427	0.144	519342933	0.461
0.125X_6DAS	54060806	0.304	513832505	0.081	519342933	0.268
0.125X_0.5X	54061791	0.150	513831381	0.505	519342933	0.083
n	6		4		1	
Bonf	0.008		0.013		0.050	

Table 3.11. Continued

	AT5G52840		AT5G64280		AT5G64290	
	SNP	P	SNP	P	SNP	P
Grand Mean	521413701	0.056	525711093	0.721	525714963	0.341
Mean_14DAS	521413701	0.227	525711093	0.428	525715294	0.861
Mean_6DAS	521413701	0.429	525711093	0.905	525715294	0.412
Mean_0.5X	521413701	0.068	525711093	0.324	525714963	0.116
0.25X_14DAS	521414157	0.126	525711093	0.468	525714963	0.633
0.25X_6DAS	521413701	0.155	525711093	0.638	525714963	0.894
0.25X_0.5X	521413701	0.147	525711093	0.841	525714963	0.025[a]
0.125X_14DAS	521413701	0.048[a]	525711093	0.748	525716073	0.439
0.125X_6DAS	521414318	0.237	525711093	0.384	525714963	0.326
0.125X_0.5X	521413701	0.102	525711093	0.288	525714963	0.598
n	4		1		3	
Bonf	0.013		0.050		0.017	

Table 3.12. The top SNPs within and 10kb up and downstream the genes involved in photorespiration for each phenotype from the EMMAX model using 211K SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. The subscript 'b' indicates SNPs that are significant with the Bonferroni correction.

	AT1G11860		**AT1G14450**		**AT1G23310**	
	SNP	P	SNP	P	SNP	P
Grand Mean	14006697	0.051	14955127	0.071	18280452	0.036[a]
Mean_14DAS	14008331	0.049[a]	14947128	0.037[a]	18275622	0.073
Mean_6DAS	13999019	0.090	14954192	0.023[a]	18275528	0.099
Mean_0.5X	14006697	0.011[a]	14955127	0.014[a]	18280452	0.074
0.25X_14DAS	14001906	0.117	14946664	0.047[a]	18275622	0.058
0.25X_6DAS	14001906	0.040[a]	14953578	0.015[a]	18275528	0.012[a]
0.25X_0.5X	14001906	0.023[a]	14955127	0.005[a]	18264642	0.020[a]
0.125X_14DAS	14006697	0.016[a]	14947128	0.035[a]	18280452	0.003[a]
0.125X_6DAS	13999106	0.041[a]	14954192	0.022[a]	18280452	0.164
0.125X_0.5X	14006697	0.018[a]	14937255	0.029[a]	18262822	0.022[a]
n	75		26		51	
Bonf	0.001		0.002		0.001	
	AT1G32470		**AT1G48030**		**AT1G67350**	
	SNP	P	SNP	P	SNP	P
Grand Mean	111740236	0.087	117716304	0.008[a]	125237293	0.067
Mean_14DAS	111745033	0.037[a]	117720042	0.025[a]	125237293	0.017[a]
Mean_6DAS	111749241	0.064	117720415	0.037[a]	125237108	0.038[a]
Mean_0.5X	111734781	0.014[a]	117720042	0.005[a]	125236635	0.009[a]
0.25X_14DAS	111745033	0.044[a]	117713326	0.023[a]	125230542	0.047[a]
0.25X_6DAS	111749074	0.122	117720415	0.023[a]	125230542	0.047[a]
0.25X_0.5X	111749852	0.019[a]	117720415	0.013[a]	125236635	0.010[a]
0.125X_14DAS	111740236	0.053	117720042	0.033[a]	125237108	0.032[a]
0.125X_6DAS	111749241	0.082	117716304	0.003[a]	125245461	0.015[a]
0.125X_0.5X	111742930	0.044[a]	117716304	0.007[a]	125232343	0.031[a]
n	32		52		33	
Bonf	0.002		0.001		0.002	

Table 3.12 Continued.

	AT1G68010		**AT1G70580**		**AT1G80380**	
	SNP	P	SNP	P	SNP	P
Grand Mean	125504497	0.020[a]	126617288	0.007[a]	130227036	0.007[a]
Mean_14DAS	125497369	0.004[a]	126619364	0.068[a]	130214813	0.010[a]
Mean_6DAS	125500681	0.079	126619364	0.039[a]	130220072	0.139
Mean_0.5X	125504497	0.001[a]	126617288	0.005[a]	130227036	0.006[a]
0.25X_14DAS	125497369	0.012[a]	126617288	0.049[a]	130227036	0.001[a]
0.25X_6DAS	125488273	0.024[a]	126617288	0.052	130224225	0.123
0.25X_0.5X	125504497	0.019[a]	126617288	0.007[a]	130227036	0.003[a]
0.125X_14DAS	125504497	0.044[a]	126619364	0.043[a]	130216388	0.057
0.125X_6DAS	125483618	0.096	126619364	0.033[a]	130217054	0.045[a]
0.125X_0.5X	125504497	0.013[a]	126617288	0.041[a]	130216388	0.073
n	32		15		50	
Bonf	0.002		0.003		0.001	
	AT2G02050		**AT2G04540**		**AT2G13360**	
	SNP	P	SNP	P	SNP	P
Grand Mean	2485656	0.161	21590365	0.018[a]	25542392	0.112
Mean_14DAS	2494321	0.034[a]	21584422	0.040[a]	25551342	0.077
Mean_6DAS	2485656	0.126	21590365	0.020[a]	25540226	0.039[a]
Mean_0.5X	2497549	0.032[a]	21581526	0.046[a]	25540646	0.105
0.25X_14DAS	2494321	0.007[a]	21584422	0.016[a]	25545672	0.024[a]
0.25X_6DAS	2498702	0.234	21585552	0.040[a]	25542733	0.008[a]
0.25X_0.5X	2497458	0.029[a]	21586375	0.015[a]	25542392	0.086
0.125X_14DAS	2486627	0.083	21581526	0.030[a]	25549098	0.056
0.125X_6DAS	2494189	0.139	21590365	0.015[a]	25540226	0.009[a]
0.125X_0.5X	2497549	0.002[a]	21581526	0.009[a]	25540646	0.119
n	36		56		35	
Bonf	0.001		0.001		0.001	

Table 3.12. Continued

	AT2G26080		**AT2G27730**		**AT2G33220**	
	SNP	P	SNP	P	SNP	P
Grand Mean	211099054	0.008[a]	211821683	0.066	214074584	0.072
Mean_14DAS	211099054	0.004[a]	211821683	0.040[a]	214081348	0.015[a]
Mean_6DAS	211099054	0.036[a]	211825246	0.020[a]	214070416	0.083
Mean_0.5X	211114928	0.007[a]	211821683	0.040[a]	214081348	0.010[a]
0.25X_14DAS	211099054	0.003[a]	211821683	0.063	214081348	0.005[a]
0.25X_6DAS	211099054	0.010[a]	211812904	0.112	214070416	0.349
0.25X_0.5X	211109872	0.007[a]	211813102	0.131	214081894	0.006[a]
0.125X_14DAS	211119393	0.006[a]	211826809	0.080	214074623	0.227
0.125X_6DAS	211119393	0.065	211823002	0.020[a]	214074584	0.044[a]
0.125X_0.5X	211114894	0.036[a]	211821683	0.056	214081348	0.125
n	36		25		24	
Bonf	0.001		0.002		0.002	
	AT2G35120		**AT2G35370**		**AT2G41220**	
	SNP	P	SNP	P	SNP	P
Grand Mean	214805846	0.072	214900180	0.011[a]	217186417	0.036[a]
Mean_14DAS	214803771	0.081	214900180	0.012[a]	217186417	0.093
Mean_6DAS	214807975	0.046[a]	214900569	0.038[a]	217186417	0.032[a]
Mean_0.5X	214805846	0.097	214882252	0.032[a]	217186417	0.025[a]
0.25X_14DAS	214803771	0.046[a]	214900180	0.002[a]	217186417	0.208
0.25X_6DAS	214807975	0.102	214882252	0.070	217170904	0.154
0.25X_0.5X	214814558	0.024[a]	214882252	0.037[a]	217186417	0.203
0.125X_14DAS	214812348	0.060	214900003	0.034[a]	217177677	0.151
0.125X_6DAS	214814558	0.030[a]	214900569	0.044[a]	217186417	0.014[a]
0.125X_0.5X	214805846	0.028[a]	214900180	0.009[a]	217186417	0.020[a]
n	20		31		12	
Bonf	0.003		0.002		0.004	

Table 3.12. Continued

	AT2G47690		**AT3G14415**		**AT3G14420**	
	SNP	P	SNP	P	SNP	P
Grand Mean	219542811	0.002[a]	34815256	0.017[a]	34833426	0.317
Mean_14DAS	219542811	0.007[a]	34815256	0.004[a]	34833426	0.639
Mean_6DAS	219542811	0.026[a]	34815256	0.013[a]	34833426	0.460
Mean_0.5X	219542811	0.008[a]	34827454	0.061	34832020	0.154
0.25X_14DAS	219542811	0.001[a]	34815256	0.039[a]	34832020	0.820
0.25X_6DAS	219542811	0.007[a]	34812265	0.091	34833426	0.616
0.25X_0.5X	219542811	0.005[a]	34812780	0.009[a]	34832020	0.916
0.125X_14DAS	219547334	0.117	34815256	0.007[a]	34833426	0.217
0.125X_6DAS	219560800	0.072	34815256	0.043[a]	34833426	0.379
0.125X_0.5X	219550445	0.006[a]	34829797	0.044[a]	34832020	0.038[a]
n	42		29		2	
Bonf	0.001		0.002		0.025	
	AT3G17240		**AT3G18410**		**AT3G54110**	
	SNP	P	SNP	P	SNP	P
Grand Mean	35899140	0.038[a]	36318100	0.014[a]	320048548	0.045[a]
Mean_14DAS	35899140	0.015[a]	36318100	0.162	320048548	0.014[a]
Mean_6DAS	35889399	0.004[a]	36330925	0.016	320048548	0.073
Mean_0.5X	35895677	0.029[a]	36323815	0.050	320050953	0.069
0.25X_14DAS	35884309	0.001[b]	36330727	0.106	320031133	0.061
0.25X_6DAS	35889399	0.026[a]	36330925	0.040[a]	320048548	0.069
0.25X_0.5X	35900452	0.008[a]	36326530	0.020	320045124	0.053
0.125X_14DAS	35882620	0.012[a]	36318100	0.121	320048548	0.011[a]
0.125X_6DAS	35889399	0.005[a]	36330925	0.019[a]	320038156	0.064
0.125X_0.5X	35880477	0.083	36318100	0.075	320034944	0.022[a]
n	53		36		47	
Bonf	0.001		0.001		0.001	

Table 3.12. Continued

	AT4G16450		**AT4G33010**		**AT4G35090**	
	SNP	P	SNP	P	SNP	P
Grand Mean	49280248	0.007[a]	415921911	0.085	416711934	0.014[a]
Mean_14DAS	49272335	0.019[a]	415930433	0.099	416710328	0.020[a]
Mean_6DAS	49280248	0.010[a]	415930433	0.021[a]	416694756	0.011[a]
Mean_0.5X	49280248	0.069	415918661	0.001[b]	416703032	0.005[a]
0.25X_14DAS	49272335	0.062	415921911	0.200	416698753	0.049[a]
0.25X_6DAS	49271609	0.041[a]	415930433	0.032[a]	416700991	0.040[a]
0.25X_0.5X	49286029	0.125	415929822	0.002[b]	416698753	0.025[a]
0.125X_14DAS	49280248	0.013[a]	415918661	0.017[a]	416710328	0.015[a]
0.125X_6DAS	49280248	0.007[a]	415926353	0.029[a]	416709405	0.018[a]
0.125X_0.5X	49280248	0.047[a]	415918661	0.044[a]	416703032	0.005[a]
n	57		27		49	
Bonf	0.001		0.002		0.001	
	AT4G37930		**AT5G04140**		**AT5G06580**	
	SNP	P	SNP	P	SNP	P
Grand Mean	417823919	0.027[a]	51132202	0.004[a]	52024278	0.090
Mean_14DAS	417837235	0.104	51132202	0.007[a]	52024278	0.061
Mean_6DAS	417823919	0.006[a]	51132202	0.017[a]	52024278	0.120
Mean_0.5X	417835086	0.140	51132202	0.041[a]	52014746	0.056
0.25X_14DAS	417823125	0.079	51132202	0.005[a]	52004060	0.151
0.25X_6DAS	417823919	0.003[a]	51140210	0.079	52024278	0.134
0.25X_0.5X	417823125	0.024[a]	51132202	0.047[a]	52024278	0.110
0.125X_14DAS	417822597	0.040[a]	51146286	0.078	52024278	0.066
0.125X_6DAS	417837235	0.011[a]	51132202	0.017[a]	52022663	0.080
0.125X_0.5X	417830131	0.174	51140210	0.027[a]	52014249	0.045[a]
n	19		28		18	
Bonf	0.003		0.002		0.003	

Table 3.12. Continued

	AT5G12860		**AT5G35630**		**AT5G47760**	
	SNP	P	SNP	P	SNP	P
Grand Mean	54062895	0.024[a]	513822767	0.006[a]	519339748	0.024[a]
Mean_14DAS	54050817	0.007[a]	513823553	0.007[a]	519339264	0.055
Mean_6DAS	54067442	0.018[a]	513822087	0.010[a]	519339748	0.015[a]
Mean_0.5X	54050817	0.001[b]	513822767	0.021[a]	519354663	0.011[a]
0.25X_14DAS	54050817	0.017[a]	513823261	0.007[a]	519338022	0.110
0.25X_6DAS	54067442	0.014[a]	513822087	0.005[a]	519339264	0.006[a]
0.25X_0.5X	54050817	0.020[a]	513822767	0.026[a]	519351728	0.008[a]
0.125X_14DAS	54062895	0.003[a]	513823553	0.003[a]	519338022	0.127
0.125X_6DAS	54062895	0.074	513823553	0.013[a]	519339748	0.072
0.125X_0.5X	54050817	0.005[a]	513822767	0.062	519335976	0.044[a]
n	35		35		47	
Bonf	0.001		0.001		0.001	
	AT5G52840		**AT5G64280**		**AT5G64290**	
	SNP	P	SNP	P	SNP	P
Grand Mean	521422508	0.001[b]	525705167	0.190	525725172	0.286
Mean_14DAS	521422508	0.002[a]	525718280	0.125	525725346	0.231
Mean_6DAS	521407946	0.000[b]	525705167	0.107	525726249	0.407
Mean_0.5X	521423609	0.020[a]	525707142	0.006[a]	525725172	0.072
0.25X_14DAS	521413626	0.012[a]	525717436	0.041[a]	525723760	0.280
0.25X_6DAS	521416728	0.007[a]	525721202	0.072	525726249	0.414
0.25X_0.5X	521416728	0.063	525718360	0.010[a]	525725346	0.053
0.125X_14DAS	521422508	0.000[b]	525702570	0.260	525725346	0.359
0.125X_6DAS	521422508	0.000[b]	525702570	0.199	525725259	0.629
0.125X_0.5X	521423609	0.029[a]	525707142	0.036[a]	525726249	0.483
n	43		33		5	
Bonf	0.001		0.002		0.01	

Table 3.13. The top SNPs within the genes involved in photorespiration for each phenotype from the EMMAX model using 1.6M SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. The subscript 'b' indicates SNPs that are significant with the Bonferroni correction.

	AT1G11860		**AT1G14450**		**AT1G23310**	
	SNP	P	SNP	P	SNP	P
Grand Mean	14001906	0.151	14946225	0.041[a]	18268422	0.045[a]
Mean_6DAS	14001626	0.183	14946225	0.006[a]	18268422	0.197
Mean_14DAS	14001468	0.059	14947082	0.011[a]	18269131	0.131
Mean_0.5X	14003257	0.048[a]	14946664	0.213	18268422	0.262
0.25X_0.5X	14001906	0.022[a]	14946664	0.316	18268422	0.085
0.25X_6DAS	14001906	0.025[a]	14946225	0.008[a]	18269835	0.020[a]
0.25X_14DAS	14001906	0.140	14946225	0.021[a]	18268422	0.069
0.125X_0.5X	14003257	0.031[a]	14947328	0.480	18269549	0.212
0.125X_6DAS	14002963	0.072	14946225	0.053	18269062	0.346
0.125X_14DAS	14002668	0.046[a]	14947128	0.005[a]	18269549	0.044[a]
n	34		11		10	
Bonf	0.001		0.005		0.005	
	AT1G32470		**AT1G48030**		**AT1G67350**	
	SNP	P	SNP	P	SNP	P
Grand Mean	111739669	0.121	117717459	0.030[a]	125236635	0.024[a]
Mean_6DAS	111739738	0.229	117717630	0.116	125236226	0.046[a]
Mean_14DAS	111739626	0.022[a]	117717459	0.158	125235584	0.200
Mean_0.5X	111739497	0.064	117717630	0.039[a]	125236635	0.003[a]
0.25X_0.5X	111739497	0.182	117717459	0.014[a]	125236635	0.006[a]
0.25X_6DAS	111739678	0.568	117717459	0.108	125236605	0.094
0.25X_14DAS	111739626	0.146	117717459	0.050	125236710	0.035[a]
0.125X_0.5X	111739497	0.122	117717630	0.071	125236872	0.027[a]
0.125X_6DAS	111739738	0.231	117717630	0.020[a]	125236226	0.008[a]
0.125X_14DAS	111739626	0.036[a]	117717630	0.077	125236226	0.169
n	13		5		54	
Bonf	0.004		0.010		0.001	

Table 3.13. Continued

	AT1G68010		AT1G70580		AT1G80380	
	SNP	P	SNP	P	SNP	P
Grand Mean	125494350	0.137	126614090	0.097	130217235	0.135
Mean_6DAS	125493464	0.113	126614090	0.044[a]	130219633	0.050
Mean_14DAS	125494350	0.024[a]	126615762	0.089	130217346	0.035[a]
Mean_0.5X	125495052	0.235	126615377	0.062	130217235	0.049[a]
0.25X_0.5X	125495052	0.132	126614090	0.010[a]	130217235	0.020[a]
0.25X_6DAS	125495208	0.166	126614090	0.012[a]	130219633	0.204
0.25X_14DAS	125494350	0.095	126614090	0.044	130217346	0.059
0.125X_0.5X	125495289	0.160	126615377	0.211	130219609	0.141
0.125X_6DAS	125493464	0.130	126614090	0.364	130217587	0.050
0.125X_14DAS	125494708	0.036[a]	126615577	0.298	130217346	0.098
n	13		8		10	
Bonf	0.004		0.006		0.005	
	AT2G02050		**AT2G04540**		**AT2G13360**	
	SNP	P	SNP	P	SNP	P
Grand Mean	2491728	0.342	21581526	0.045[a]	25541321	0.010[a]
Mean_6DAS	2490568	0.275	21584092	0.096	25541321	0.011[a]
Mean_14DAS	2491876	0.289	21581687	0.023[a]	25541321	0.009[a]
Mean_0.5X	2490824	0.033[a]	21581526	0.073	25540646	0.142
0.25X_0.5X	2491876	0.043[a]	21584553	0.088	25540933	0.231
0.25X_6DAS	2491982	0.360	21584680	0.039[a]	25540447	0.012[a]
0.25X_14DAS	2491876	0.120	21581687	0.025[a]	25541321	0.026[a]
0.125X_0.5X	2491374	0.004[b]	21581526	0.016[a]	25539741	0.135
0.125X_6DAS	2490386	0.140	21584092	0.063	25541321	0.007[a]
0.125X_14DAS	2490824	0.314	21584317	0.007[a]	25541321	0.015[a]
n	9		94		40	
Bonf	0.006		0.001		0.001	

Table 3.13. Continued

	AT2G26080		**AT2G27730**		**AT2G33220**	
	SNP	P	SNP	P	SNP	P
Grand Mean	211112112	0.028[a]	211821683	0.017[a]	214079054	0.074
Mean_6DAS	211111305	0.053	211820616	0.116	214079725	0.016[a]
Mean_14DAS	211111305	0.015[a]	211821683	0.011[a]	214079486	0.080
Mean_0.5X	211110546	0.040[a]	211821683	0.008[a]	214080087	0.023[a]
0.25X_0.5X	211111305	0.074	211821683	0.062	214078960	0.002[a]
0.25X_6DAS	211111305	0.021[a]	211820616	0.195	214079725	0.092
0.25X_14DAS	211111305	0.048[a]	211821683	0.032[a]	214080077	0.170
0.125X_0.5X	211109479	0.149	211821683	0.017[a]	214080087	0.138
0.125X_6DAS	211110738	0.042[a]	211821729	0.178	214079725	0.018[a]
0.125X_14DAS	211111305	0.040[a]	211821683	0.046[a]	214079739	0.139
n	25		10		46	
Bonf	0.002		0.005		0.001	
	AT2G35120		**AT2G35370**		**AT2G41220**	
	SNP	P	SNP	P	SNP	P
Grand Mean	214805846	0.087	214891195	0.291	217186417	0.088
Mean_6DAS	214805846	0.131	214892038	0.653	217177911	0.038[a]
Mean_14DAS	214806278	0.182	214892038	0.218	217177911	0.060
Mean_0.5X	214805846	0.077	214892038	0.524	217178309	0.031[a]
0.25X_0.5X	214807148	0.083	214892038	0.570	217185955	0.159
0.25X_6DAS	214805846	0.241	214892038	0.239	217180156	0.268
0.25X_14DAS	214806048	0.187	214892038	0.271	217185955	0.051
0.125X_0.5X	214805846	0.019[a]	214891195	0.134	217178309	0.021[a]
0.125X_6DAS	214807148	0.115	214891195	0.801	217178309	0.021[a]
0.125X_14DAS	214805846	0.234	214892038	0.330	217177911	0.041[a]
n	11		3		15	
Bonf	0.005		0.017		0.003	

Table 3.13. Continued

	AT2G47690		**AT3G14415**		**AT3G14420**	
	SNP	P	SNP	P	SNP	P
Grand Mean	219552346	0.262	34820182	0.211	34823237	0.422
Mean_6DAS	219552346	0.853	34818332	0.178	34823237	0.328
Mean_14DAS	219552346	0.218	34818503	0.072	34823237	0.072
Mean_0.5X	219552346	0.220	34820182	0.098	34823150	0.599
0.25X_0.5X	219552346	0.193	34819102	0.015[a]	34822602	0.210
0.25X_6DAS	219552346	0.846	34818332	0.219	34822497	0.121
0.25X_14DAS	219552346	0.302	34818503	0.072	34823237	0.072
0.125X_0.5X	219552346	0.360	34818631	0.373	34823150	0.158
0.125X_6DAS	219552346	0.902	34818503	0.242	34823237	0.242
0.125X_14DAS	219552346	0.209	34818332	0.116	34822929	0.103
n	2		13		5	
Bonf	0.025		0.004		0.010	
	AT3G17240		**AT3G18410**		**AT3G54110**	
	SNP	P	SNP	P	SNP	P
Grand Mean	35890668	0.015[a]	36324384	0.425	320040146	0.071
Mean_6DAS	35891602	0.103	36324050	0.327	320039392	0.153
Mean_14DAS	35892091	0.077	36324384	0.361	320040146	0.175
Mean_0.5X	35890668	0.029[a]	36323815	0.043[a]	320041003	0.107
0.25X_0.5X	35890668	0.026[a]	36323815	0.130	320039982	0.164
0.25X_6DAS	35891602	0.059	36324363	0.409	320040146	0.196
0.25X_14DAS	35890668	0.005[a]	36323180	0.402	320040146	0.049[a]
0.125X_0.5X	35891549	0.092	36324208	0.106	320041003	0.040[a]
0.125X_6DAS	35891340	0.074	36324050	0.114	320039392	0.140
0.125X_14DAS	35892014	0.014	36324051	0.197	320039917	0.120
n	22		8		38	
Bonf	0.002		0.006		0.001	

Table 3.13. Continued

	AT4G16450		AT4G33010		AT4G35090	
	SNP	P	SNP	P	SNP	P
Grand Mean	49280248	0.012[a]	415930995	0.060	416700726	0.025[a]
Mean_6DAS	49280248	0.015[a]	415930995	0.039[a]	416700814	0.025[a]
Mean_14DAS	49280248	0.177	415927552	0.219	416700814	0.018[a]
Mean_0.5X	49280742	0.004[b]	415930995	0.037[a]	416703032	0.003[a]
0.25X_0.5X	49280742	0.011[a]	415929822	0.000[b]	416702962	0.036[a]
0.25X_6DAS	49280742	0.075	415930995	0.051	416700814	0.014[a]
0.25X_14DAS	49280248	0.426	415930995	0.469	416700435	0.048[a]
0.125X_0.5X	49280248	0.019[a]	415927833	0.186	416703032	0.002[a]
0.125X_6DAS	49280248	0.005[a]	415930995	0.057	416700726	0.010[a]
0.125X_14DAS	49280248	0.028[a]	415927552	0.098	416701493	0.010[a]
n	10		15		58	
Bonf	0.005		0.003		0.001	
	AT4G37930		**AT5G04140**		**AT5G06580**	
	SNP	P	SNP	P	SNP	P
Grand Mean	417832670	0.028[a]	51132202	0.025[a]	52011963	0.019[a]
Mean_6DAS	417833773	0.011[a]	51132202	0.039[a]	52011963	0.007[a]
Mean_14DAS	417832758	0.051	51135175	0.025[a]	52011963	0.057
Mean_0.5X	417834603	0.020[a]	51132202	0.132	52011963	0.090
0.25X_0.5X	417833773	0.010[a]	51135104	0.029[a]	52014746	0.144
0.25X_6DAS	417833773	0.001[a]	51131072	0.087	52011963	0.044[a]
0.25X_14DAS	417832758	0.023[a]	51135175	0.027[a]	52011905	0.096
0.125X_0.5X	417834603	0.086	51130693	0.205	52013686	0.015[a]
0.125X_6DAS	417832670	0.083	51130377	0.018[a]	52011963	0.037[a]
0.125X_14DAS	417832670	0.133	51130367	0.177	52011963	0.054
n	74		43		13	
Bonf	0.001		0.001		0.004	

Table 3.13. Continued

	AT5G12860		AT5G35630		AT5G47760	
	SNP	P	SNP	P	SNP	P
Grand Mean	54061791	0.083	513831196	0.034[a]	519342933	0.081
Mean_6DAS	54060296	0.116	513832505	0.045[a]	519342985	0.075
Mean_14DAS	54061791	0.026[a]	513830606	0.008[a]	519342985	0.168
Mean_0.5X	54061791	0.093	513830640	0.021[a]	519342933	0.009[a]
0.25X_0.5X	54061875	0.021[a]	513830640	0.038[a]	519344461	0.010[a]
0.25X_6DAS	54060296	0.091	513831196	0.066	519342985	0.087
0.25X_14DAS	54061791	0.072	513830606	0.012[a]	519342985	0.193
0.125X_0.5X	54061791	0.124	513830640	0.231	519342933	0.049
0.125X_6DAS	54060296	0.224	513832505	0.033[a]	519342993	0.156
0.125X_14DAS	54061791	0.033[a]	513831196	0.103	519342993	0.464
n	9		39		32	
Bonf	0.006		0.001		0.002	
	AT5G52840		**AT5G64280**		**AT5G64290**	
	SNP	P	SNP	P	SNP	P
Grand Mean	521413701	0.061	525711679	0.079	525715171	0.238
Mean_6DAS	521414318	0.500	525711679	0.028[a]	525716073	0.170
Mean_14DAS	521414184	0.217	525711243	0.109	525716661	0.168
Mean_0.5X	521413701	0.052	525713233	0.073	525715171	0.051
0.25X_0.5X	521413701	0.145	525712588	0.020[a]	525714963	0.020[a]
0.25X_6DAS	521413701	0.259	525711679	0.137	525715734	0.310
0.25X_14DAS	521414184	0.168	525711243	0.079	525715703	0.190
0.125X_0.5X	521413701	0.051	525713233	0.148	525715693	0.243
0.125X_6DAS	521414157	0.135	525711679	0.080	525714963	0.189
0.125X_14DAS	521413701	0.031[a]	525712588	0.055	525715693	0.126
n	6		11		37	
Bonf	0.008		0.005		0.001	

Table 3.14. The top SNPs within and 10kb up and downstream the genes involved in photorespiration for each phenotype from the EMMAX model using 1.6M SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. The subscript 'b' indicates SNPs that are significant with the Bonferroni correction.

	AT1G11860		**AT1G14450**		**AT1G23310**	
	SNP	P	SNP	P	SNP	P
Grand Mean	14003571	0.019[a]	14941620	0.031[a]	18274103	0.001[a]
Mean_6DAS	14010792	0.043[a]	14946225	0.006[a]	18274103	0.001[a]
Mean_14DAS	14007800	0.021[a]	14948420	0.006[a]	18274103	0.006[a]
Mean_0.5X	14008301	0.005[a]	14955127	0.009[a]	18278631	0.016[a]
0.25X_0.5X	14006862	0.011[a]	14955127	0.003[a]	18278815	0.003[a]
0.25X_6DAS	14008064	0.018[a]	14953738	0.007[a]	18277051	0.000[a]
0.25X_14DAS	14010145	0.056	14946225	0.021[a]	18274103	0.003[a]
0.125X_0.5X	14007078	0.001[a]	14949081	0.050[a]	18267041	0.011[a]
0.125X_6DAS	14007800	0.003[a]	14941620	0.016[a]	18274103	0.008[a]
0.125X_14DAS	14007068	0.002[a]	14947128	0.005[a]	18280452	0.004[a]
	444		150		349	
Bonf	0.000		0.000		0.000	
	AT1G32470		**AT1G48030**		**AT1G67350**	
	SNP	P	SNP	P	SNP	P
Grand Mean	111732357	0.031[a]	117710508	0.000[a]	125233365	0.018[a]
Mean_6DAS	111748233	0.009[a]	117727585	0.001[a]	125233564	0.008[a]
Mean_14DAS	111745033	0.007[a]	117710508	0.000[b]	125233085	0.027[a]
Mean_0.5X	111732357	0.004[a]	117716304	0.005[a]	125233420	0.002[a]
0.25X_0.5X	111741323	0.020[a]	117716816	0.007[a]	125236635	0.006[a]
0.25X_6DAS	111737668	0.046[a]	117713452	0.002[a]	125227453	0.004[a]
0.25X_14DAS	111745033	0.011[a]	117710508	0.002[a]	125233085	0.007[a]
0.125X_0.5X	111732100	0.012[a]	117727141	0.003[a]	125233420	0.008[a]
0.125X_6DAS	111748233	0.008[a]	117710426	0.001[a]	125234442	0.007[a]
0.125X_14DAS	111741589	0.011[a]	117710508	0.000[a]	125227127	0.074
	241		378		311	
Bonf	0.000		0.000		0.000	

Table 3. 14. Continued

	AT1G68010		AT1G70580		AT1G80380	
	SNP	P	SNP	P	SNP	P
Grand Mean	125496226	0.009[a]	126603294	0.004[a]	130227036	0.016[a]
Mean_6DAS	125496226	0.009[a]	126619588	0.002[a]	130214998	0.014[a]
Mean_14DAS	125496226	0.001[a]	126603122	0.008[a]	130227972	0.006[a]
Mean_0.5X	125504497	0.018[a]	126603621	0.007[a]	130229399	0.005[a]
0.25X_0.5X	125496226	0.012[a]	126617736	0.007[a]	130227036	0.002[a]
0.25X_6DAS	125496226	0.009[a]	126625091	0.004[a]	130212339	0.014[a]
0.25X_14DAS	125496226	0.000[a]	126603122	0.020[a]	130227972	0.002[a]
0.125X_0.5X	125502083	0.013[a]	126604215	0.024[a]	130229399	0.018[a]
0.125X_6DAS	125487784	0.002[a]	126619588	0.004[a]	130227589	0.006[a]
0.125X_14DAS	125486729	0.018[a]	126618568	0.034[a]	130227589	0.034[a]
	278		109		204	
Bonf	0.000		0.000		0.000	
	AT2G02050		**AT2G04540**		**AT2G13360**	
	SNP	P	SNP	P	SNP	P
Grand Mean	2496515	0.013[a]	21593028	0.002[a]	25542449	0.002[a]
Mean_6DAS	2496515	0.012[a]	21586641	0.002[a]	25542449	0.003[a]
Mean_14DAS	2494377	0.007[a]	21593120	0.011[a]	25541321	0.009[a]
Mean_0.5X	2485409	0.007[a]	21593028	0.000[a]	25530931	0.017[a]
0.25X_0.5X	2483330	0.007[a]	21593028	0.000[a]	25533671	0.002[a]
0.25X_6DAS	2496515	0.004[a]	21588009	0.004[a]	25542733	0.011[a]
0.25X_14DAS	2494377	0.003[a]	21593120	0.007[a]	25536270	0.024[a]
0.125X_0.5X	2493516	0.001[a]	21587399	0.003[a]	25529883	0.002[a]
0.125X_6DAS	2485783	0.003[a]	21586641	0.002[a]	25541321	0.007[a]
0.125X_14DAS	2485598	0.009[a]	21587399	0.001[a]	25541321	0.015[a]
	340		516		314	
Bonf	0.000		0.000		0.000	

Table 3. 14. Continued

	AT2G26080		AT2G27730		AT2G33220	
	SNP	P	SNP	P	SNP	P
Grand Mean	211123263	0.003[a]	211821683	0.017[a]	214070175	0.061
Mean_6DAS	211123263	0.000[a]	211826675	0.000[b]	214079725	0.016[a]
Mean_14DAS	211099054	0.002[a]	211818853	0.005[a]	214080227	0.008[a]
Mean_0.5X	211120288	0.003[a]	211821683	0.008[a]	214074584	0.022[a]
0.25X_0.5X	211108491	0.017[a]	211809981	0.017[a]	214078960	0.002[a]
0.25X_6DAS	211123650	0.003[a]	211826675	0.002[a]	214078305	0.032[a]
0.25X_14DAS	211099054	0.002[a]	211828917	0.005[a]	214080227	0.002[a]
0.125X_0.5X	211099924	0.012[a]	211821683	0.017[a]	214078698	0.031[a]
0.125X_6DAS	211103281	0.002[a]	211822072	0.000[a]	214070175	0.017[a]
0.125X_14DAS	211123263	0.002[a]	211814093	0.010[a]	214078297	0.023[a]
	258		239		220	
Bonf	0.000		0.000		0.000	
	AT2G35120		**AT2G35370**		**AT2G41220**	
	SNP	P	SNP	P	SNP	P
Grand Mean	214804163	0.009[a]	214901323	0.006[a]	217190282	0.011[a]
Mean_6DAS	214811667	0.007[a]	214892567	0.035[a]	217177911	0.038[a]
Mean_14DAS	214813630	0.010[a]	214892567	0.005[a]	217190282	0.035[a]
Mean_0.5X	214813105	0.011[a]	214894088	0.013[a]	217178309	0.031[a]
0.25X_0.5X	214804239	0.014[a]	214887710	0.029[a]	217190006	0.005[a]
0.25X_6DAS	214812815	0.008[a]	214892567	0.031[a]	217190282	0.029[a]
0.25X_14DAS	214812898	0.005[a]	214892567	0.001[a]	217190282	0.040[a]
0.125X_0.5X	214810792	0.009[a]	214894657	0.003[a]	217178309	0.021[a]
0.125X_6DAS	214811667	0.007[a]	214883910	0.056	217177612	0.017[a]
0.125X_14DAS	214813630	0.009[a]	214894657	0.025[a]	217195804	0.025[a]
	185		179		93	
Bonf	0.000		0.000		0.001	

Table 3. 14. Continued

	AT2G47690		AT3G14415		AT3G14420	
	SNP	P	SNP	P	SNP	P
Grand Mean	219542811	0.001[a]	34821302	0.017[a]	34833045	0.156
Mean_6DAS	219542811	0.007[a]	34814015	0.024[a]	34833045	0.217
Mean_14DAS	219542811	0.007[a]	34821302	0.020[a]	34833956	0.045[a]
Mean_0.5X	219542811	0.012[a]	34809932	0.004[a]	34832928	0.039[a]
0.25X_0.5X	219541515	0.008[a]	34809932	0.002[a]	34833230	0.367
0.25X_6DAS	219542811	0.001[a]	34814015	0.022[a]	34833045	0.033[a]
0.25X_14DAS	219542811	0.000[a]	34821302	0.003[a]	34833956	0.040[a]
0.125X_0.5X	219553426	0.002[a]	34826659	0.004[a]	34832928	0.015[a]
0.125X_6DAS	219560800	0.034[a]	34827562	0.030[a]	34833230	0.187
0.125X_14DAS	219543134	0.017[a]	34808466	0.031[a]	34833426	0.320
	145		168		18	
Bonf	0.000		0.000		0.003	
	AT3G17240		**AT3G18410**		**AT3G54110**	
	SNP	P	SNP	P	SNP	P
Grand Mean	35890668	0.015[a]	36318100	0.011[a]	320032247	0.025[a]
Mean_6DAS	35889399	0.001[a]	36329728	0.004[a]	320050799	0.024[a]
Mean_14DAS	35887004	0.005[a]	36331036	0.007[a]	320032247	0.019[a]
Mean_0.5X	35890668	0.029[a]	36327485	0.038[a]	320037668	0.046[a]
0.25X_0.5X	35898731	0.005[a]	36333552	0.002[a]	320047687	0.012[a]
0.25X_6DAS	35889399	0.011[a]	36329728	0.014[a]	320032247	0.009[a]
0.25X_14DAS	35881797	0.004[a]	36331036	0.051	320032247	0.003[a]
0.125X_0.5X	35880477	0.035[a]	36322750	0.036[a]	320034944	0.008[a]
0.125X_6DAS	35889399	0.002[a]	36331091	0.013[a]	320041611	0.008[a]
0.125X_14DAS	35895390	0.004[a]	36331036	0.004[a]	320047687	0.009[a]
	315		219		234	
Bonf	0.000		0.000		0.000	

Table 3. 14. Continued

	AT4G16450		AT4G33010		AT4G35090	
	SNP	P	SNP	P	SNP	P
Grand Mean	49271703	0.011[a]	415933763	0.009[a]	416691928	0.005[a]
Mean_6DAS	49280248	0.015[a]	415931796	0.001[a]	416691928	0.001[a]
Mean_14DAS	49271703	0.018[a]	415931796	0.032[a]	416699056	0.017[a]
Mean_0.5X	49280895	0.004[a]	415933763	0.000[a]	416706574	0.001[a]
0.25X_0.5X	49280895	0.002[a]	415917572	0.000[b]	416711934	0.014[a]
0.25X_6DAS	49280875	0.040[a]	415933763	0.012[a]	416691928	0.002[a]
0.25X_14DAS	49271703	0.033[a]	415921046	0.014[a]	416705553	0.009[a]
0.125X_0.5X	49280248	0.019[a]	415933763	0.013[a]	416695097	0.001[a]
0.125X_6DAS	49271703	0.001[a]	415931796	0.013[a]	416700726	0.010[a]
0.125X_14DAS	49271703	0.026[a]	415931796	0.004[a]	416701493	0.010[a]
	265		136		304	
Bonf	0.000		0.000		0.000	
	AT4G37930		**AT5G04140**		**AT5G06580**	
	SNP	P	SNP	P	SNP	P
Grand Mean	417838373	0.017[a]	51132202	0.025[a]	52011963	0.019[a]
Mean_6DAS	417823919	0.008[a]	51132202	0.039[a]	52011963	0.007[a]
Mean_14DAS	417837703	0.028[a]	51128052	0.011[a]	52007399	0.025[a]
Mean_0.5X	417837674	0.003[a]	51140412	0.039[a]	52007907	0.061
0.25X_0.5X	417830575	0.005[a]	51140412	0.007[a]	52019220	0.062
0.25X_6DAS	417833773	0.001[a]	51125041	0.032[a]	52011963	0.044[a]
0.25X_14DAS	417837703	0.010[a]	51128052	0.002[a]	52007399	0.027[a]
0.125X_0.5X	417842527	0.033[a]	51139548	0.002[a]	52007907	0.004[a]
0.125X_6DAS	417837235	0.006[a]	51130377	0.018[a]	52009649	0.030[a]
0.125X_14DAS	417824664	0.029[a]	51147114	0.036[a]	52011963	0.054
	288		189		63	
Bonf	0.000		0.000		0.001	

Table 3. 14. Continued

	AT5G12860		AT5G35630		AT5G47760	
	SNP	P	SNP	P	SNP	P
Grand Mean	54055952	0.013[a]	513823553	0.001[a]	519339716	0.015[a]
Mean_6DAS	54055952	0.009[a]	513823553	0.005[a]	519339191	0.007[a]
Mean_14DAS	54050580	0.006[a]	513823553	0.001[a]	519347947	0.012[a]
Mean_0.5X	54050817	0.000[b]	513842280	0.001[a]	519336979	0.005[a]
0.25X_0.5X	54050817	0.014[a]	513842280	0.001[a]	519347949	0.003[a]
0.25X_6DAS	54067442	0.030[a]	513822087	0.004[a]	519339264	0.015[a]
0.25X_14DAS	54050580	0.016[a]	513834271	0.003[a]	519347947	0.027[a]
0.125X_0.5X	54050580	0.004[a]	513827731	0.014[a]	519337512	0.013[a]
0.125X_6DAS	54055952	0.008[a]	513820866	0.003[a]	519336979	0.003[a]
0.125X_14DAS	54062895	0.009[a]	513823553	0.001[a]	519350732	0.078
	101		335		295	
Bonf	0.000		0.000		0.000	
	AT5G52840		**AT5G64280**		**AT5G64290**	
	SNP	P	SNP	P	SNP	P
Grand Mean	521422508	0.000[a]	525709948	0.021[a]	525726527	0.012[a]
Mean_6DAS	521422508	0.000[b]	525711679	0.028[a]	525726527	0.093
Mean_14DAS	521422508	0.001[a]	525709948	0.018[a]	525726527	0.011[a]
Mean_0.5X	521422418	0.011[a]	525707142	0.003[a]	525726527	0.006[a]
0.25X_0.5X	521408799	0.014[a]	525718360	0.011[a]	525726527	0.008[a]
0.25X_6DAS	521421814	0.003[a]	525721539	0.006[a]	525726527	0.242
0.25X_14DAS	521413626	0.012[a]	525720530	0.009[a]	525723760	0.050
0.125X_0.5X	521422566	0.014[a]	525701320	0.026[a]	525726527	0.090
0.125X_6DAS	521422508	0.000[b]	525702517	0.046[a]	525726108	0.061
0.125X_14DAS	521422418	0.000[b]	525709948	0.007[a]	525726527	0.017[a]
	224		261		19	
Bonf	0.000		0.000		0.003	

3.2.2.1.4 Putative Genes

We also tested the significance of the eight putative genes selected based on preliminary data (Table 3.1). Of these eight genes, four of the genes had SNPs within the genes that were significant for most of the phenotypes after a Bonferroni correction using the 211K SNPs dataset (Table 3.15): *ALANINE AMINOTRANSFERASE1* (*ALAAT1*), *AOP3*, *AOP1*, and *SPERMIDINE SYNTHASE3* (*SPDS3*). They also showed significance within the 20kb window surrounding the genes (Table 3.16). This is an indication that the SNPs were in linkage disequilibrium and therefore multiple SNPs had significant association with the same gene.

Two other genes, *HIGH-LEVEL EXPRESSION OF SUGAR-INDUCIBLE GENE2* (*HSI2*) and *ISOVALERYL-COA-DEHYDROGENASE* (*IVD*) showed significance in at least one phenotype within the genes and within a 20kb window of the genes using the 211K SNPs dataset (Tables 3.15 & 3.16).

AOP3, *AOP1*, and *SPDS3* also had a significant SNP within each gene for most of the phenotypes and *HSI2* was still significant in one phenotype in the 1.6M SNPs dataset (Table 3.17). A SNP within *K23L20.6*, a molybdenum cofactor sulfurase involved in metabolic processes, also was significant for the 0.25X_0.5X phenotype. Expanding the window to 10kb upstream and downstream the genes, *AOP3* and *AOP1* were still linked to significant SNPs (Table 3.18). *HSI2* and *K23L20.6* also still showed significance in at least one phenotype using the 1.6M SNPs dataset (Tables 3.17 & 3.18). Interestingly, a SNP within the 20kb window of *ALAAT1* showed significance in the 0.125X_0.5X

Table 3.15. The top SNPs within the eight candidate genes for each phenotype from the EMMAX model using 211K SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. None of the SNPs were found to be significant after the Bonferroni correction.

	ALAAT1		***HSI2***		***IVD1***	
	SNP	P	SNP	P	SNP	P
Grand Mean	15926015	0.000[b]	212983973	0.098	316622600	0.016[a]
Mean_14DAS	15926015	0.000[b]	212983973	0.022[b]	316622528	0.041[a]
Mean_6DAS	15925576	0.001[b]	212983973	0.059	316621879	0.012[a]
Mean_0.5X	15923981	0.009[a]	212983973	0.352	316622600	0.116
0.25X_14DAS	15923981	0.002[b]	212983973	0.872	316622600	0.016[a]
0.25X_6DAS	15923364	0.042[a]	212983973	0.112	316621879	0.001[b]
0.25X_0.5X	15923981	0.031[a]	212983973	0.688	316621894	0.062
0.125X_14DAS	15926015	0.000[b]	212983973	0.000[b]	316620440	0.258
0.125X_6DAS	15925576	0.000[b]	212983973	0.059	316620358	0.043[a]
0.125X_0.5X	15923981	0.020[a]	212983973	0.058	316624333	0.194
n	13		1		13	
Bonf	0.004		0.050		0.004	
	GPAT8		***AOP3***		***AOP1***	
	SNP	P	SNP	P	SNP	P
Grand Mean	4174339	0.108	41344365	0.000[b]	41358745	0.000[b]
Mean_14DAS	4174339	0.020[a]	41344365	0.001[b]	41358745	0.002[b]
Mean_6DAS	4174339	0.068	41345954	0.000[b]	41358745	0.002[b]
Mean_0.5X	4174228	0.171	41344365	0.004[b]	41358745	0.011[a]
0.25X_14DAS	4175788	0.029[a]	41344365	0.012[a]	41358545	0.002[b]
0.25X_6DAS	4174339	0.122	41345954	0.002[b]	41359167	0.001[b]
0.25X_0.5X	4174228	0.159	41345954	0.127	41358545	0.030[a]
0.125X_14DAS	4174339	0.046[a]	41345919	0.000[b]	41358745	0.000[b]
0.125X_6DAS	4174339	0.314	41345919	0.000[b]	41358745	0.000[b]
0.125X_0.5X	4175788	0.256	41344365	0.002[b]	41358745	0.002[b]
n	6		5		6	
Bonf	0.008		0.010		0.008	

Table 15. Continued

	K23L20.6		*SPDS3*	
	SNP	P	SNP	P
Grand Mean	518044675	0.042[a]	521535920	0.001[b]
Mean_14DAS	518044902	0.017[a]	521535920	0.009[b]
Mean_6DAS	518044675	0.014[a]	521535920	0.000[b]
Mean_0.5X	518044587	0.033[a]	521535920	0.051
0.25X_14DAS	518044675	0.119	521535920	0.087
0.25X_6DAS	518044587	0.004[a]	521535920	0.006[b]
0.25X_0.5X	518044587	0.019[a]	521535281	0.002[b]
0.125X_14DAS	518044902	0.027[a]	521535920	0.001[b]
0.125X_6DAS	518044104	0.144	521535920	0.000[b]
0.125X_0.5X	518044675	0.302	521535920	0.050[a]
n	15		5	
Bonf	0.003		0.010	

Table 3.16. The top SNPs within and 10kb up and downstream the eight candidate genes for each phenotype from the EMMAX model using 211K SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. None of the SNPs were found to be significant after the Bonferroni correction.

	ALAAT1		***HSI2***		***IVD1***	
	SNP	P	SNP	P	SNP	P
Grand Mean	15926015	0.000[b]	212973846	0.003[a]	316622600	0.016[a]
Mean_14DAS	15926015	0.000[b]	212973846	0.004[a]	316630830	0.022[a]
Mean_6DAS	15925576	0.001[a]	212973846	0.005[a]	316621879	0.012[a]
Mean_0.5X	15931375	0.007[a]	212989652	0.023[a]	316625094	0.065
0.25X_14DAS	15929838	0.000[b]	212973846	0.072	316628647	0.009[a]
0.25X_6DAS	15931375	0.005[a]	212973846	0.000[b]	316621879	0.001[a]
0.25X_0.5X	15928671	0.012[a]	212975898	0.007[a]	316625094	0.017[a]
0.125X_14DAS	15926015	0.000[b]	212979611	0.000[b]	316634312	0.052
0.125X_6DAS	15925576	0.000[b]	212975098	0.007[a]	316612570	0.011[a]
0.125X_0.5X	15926570	0.004[a]	212972289	0.011[a]	316630075	0.016[a]
n	54		27		53	
Bonf	0.001		0.002		0.001	
	GPAT8		***AOP3***		***AOP1***	
	SNP	P	SNP	P	SNP	P
Grand Mean	4180919	0.011[a]	41344365	0.000[b]	41358745	0.000[b]
Mean_14DAS	4172010	0.007[a]	41356197	0.001[b]	41365141	0.001[b]
Mean_6DAS	4177937	0.003[a]	41335812	0.000[b]	41365141	0.001[b]
Mean_0.5X	4172010	0.003[a]	41356083	0.001[b]	41365317	0.000[b]
0.25X_14DAS	4172010	0.025[a]	41344365	0.012[a]	41358545	0.002[a]
0.25X_6DAS	4177937	0.012[a]	41345954	0.002[a]	41359167	0.001[a]
0.25X_0.5X	4172010	0.006[a]	41356083	0.010[a]	41364978	0.015[a]
0.125X_14DAS	4177047	0.005[a]	41343290	0.000[b]	41365141	0.000[b]
0.125X_6DAS	4177937	0.002[a]	41335812	0.000[b]	41358745	0.000[b]
0.125X_0.5X	4180919	0.019[a]	41344365	0.002[a]	41362858	0.000[b]
n	65		36		42	
Bonf	0.001		0.001		0.001	

Table 3.16. Continued

	K23L20.6		*SPDS3*	
	SNP	P	SNP	P
Grand Mean	518038599	0.004[a]	521535920	0.001[a]
Mean_14DAS	518038599	0.006[a]	521535920	0.009[a]
Mean_6DAS	518041979	0.007[a]	521535920	0.000[b]
Mean_0.5X	518039912	0.001[a]	521535920	0.051
0.25X_14DAS	518038599	0.026[a]	521526994	0.072
0.25X_6DAS	518034273	0.001[a]	521525720	0.051
0.25X_0.5X	518038599	0.006[a]	521535281	0.002[a]
0.125X_14DAS	518041422	0.007[a]	521535920	0.001[a]
0.125X_6DAS	518033859	0.048[a]	521535920	0.000[b]
0.125X_0.5X	518039912	0.014[a]	521535920	0.050[a]
n	98		61	
Bonf	0.001		0.001	

Table 3.17. The top SNPs within the eight candidate genes for each phenotype from the EMMAX model using 1.6M SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. None of the SNPs were found to be significant after the Bonferroni correction.

	ALAAT1		***HSI2***		***IVD1***	
	SNP	P	SNP	P	SNP	P
Grand Mean	15926172	0.095	212980562	0.084	316622600	0.013[a]
Mean_14DAS	15922929	0.076	212980562	0.009[a]	316623138	0.026[a]
Mean_6DAS	15926172	0.098	212980562	0.035[a]	316622373	0.008[a]
Mean_0.5X	15926172	0.026[a]	212980520	0.046[a]	316620515	0.031[a]
0.25X_14DAS	15922929	0.025[a]	212985039	0.104	316621879	0.009[a]
0.25X_6DAS	15923364	0.026[a]	212985023	0.005[a]	316621879	0.00[a]
0.25X_0.5X	15926266	0.135	212980520	0.020[a]	316624909	0.003[a]
0.125X_14DAS	15923048	0.087	212983973	0.000[b]	316624466	0.004[a]
0.125X_6DAS	15923784	0.024[a]	212984224	0.023[a]	316623518	0.008[a]
0.125X_0.5X	15926172	0.013[a]	212984481	0.017[a]	316624466	0.019[a]
n	14		22		125	
Bonf	0.004		0.002		0.000	
	GPAT8		***AOP3***		***AOP1***	
	SNP	P	SNP	P	SNP	P
Grand Mean	4176756	0.004[a]	41358745	0.000[b]	41344365	0.000[b]
Mean_14DAS	4174339	0.010[a]	41359617	0.000[b]	41344365	0.000[b]
Mean_6DAS	4176756	0.008[a]	41358745	0.000[b]	41346229	0.000[b]
Mean_0.5X	4174339	0.100	41358745	0.005[a]	41344365	0.003[b]
0.25X_14DAS	4174366	0.011[a]	41359617	0.000[b]	41344365	0.006[a]
0.25X_6DAS	4176756	0.053	41359167	0.000[b]	41346229	0.000[b]
0.25X_0.5X	4175744	0.078	41359167	0.028[a]	41345545	0.073
0.125X_14DAS	4174681	0.023[a]	41358745	0.000[b]	41345919	0.000[b]
0.125X_6DAS	4176756	0.002[a]	41358745	0.000[b]	41345919	0.000[b]
0.125X_0.5X	4174200	0.120	41358745	0.001[b]	41344365	0.001[b]
n	25		37		13	
Bonf	0.002		0.001		0.004	

Table 3.17. Continued

	K23L20.6		*SPDS3*	
	SNP	P	SNP	P
Grand Mean	518043424	0.009[a]	521535920	0.000[b]
Mean_14DAS	518043340	0.010[a]	521535920	0.001[b]
Mean_6DAS	518043424	0.003[a]	521535920	0.000[b]
Mean_0.5X	518043424	0.010[a]	521535920	0.009[a]
0.25X_14DAS	518043340	0.019[a]	521534481	0.016[a]
0.25X_6DAS	518043753	0.001[a]	521535920	0.005[a]
0.25X_0.5X	518044049	0.001[b]	521535920	0.046[a]
0.125X_14DAS	518043424	0.050[a]	521535920	0.001[b]
0.125X_6DAS	518043191	0.052	521535920	0.000[b]
0.125X_0.5X	518043424	0.085	521535920	0.013[a]
n	68		22	
Bonf	0.001		0.002	

Table 3.18. The top SNPs within and 10kb up and downstream the eight candidate genes for each phenotype from the EMMAX model using 1.6M SNPs. Each gene represents two columns, the SNP column (SNP) and the *p*-value of the SNP (P). The SNP is labeled as an id, the first digit being the chromosome and the rest being the position of the SNP. The number of SNPs found within each gene is indicated in the row labeled 'n'. The subscript 'a' indicates which SNPs are significant ($\alpha <= 0.05$). To correct for multiple testing a new significant cutoff was calculated using Bonferroni model. The new significant cutoff is indicated in the last column labeled Bonf. None of the SNPs were found to be significant after the Bonferroni correction.

	ALAAT1		***HSI2***		***IVD1***	
	SNP	P	SNP	P	SNP	P
Grand Mean	15931461	0.002[a]	212973846	0.002[a]	316613709	0.002[a]
Mean_14DAS	15928655	0.002[a]	212973846	0.005[a]	316631054	0.008[a]
Mean_6DAS	15916495	0.001[a]	212973846	0.001[a]	316611994	0.001[a]
Mean_0.5X	15918198	0.003[a]	212971761	0.016[a]	316613709	0.001[a]
0.25X_14DAS	15928655	0.014[a]	212971487	0.047[a]	316613140	0.005[a]
0.25X_6DAS	15931479	0.005[a]	212973846	0.000[b]	316621879	0.001[a]
0.25X_0.5X	15918540	0.005[a]	212979085	0.005[a]	316624909	0.003[a]
0.125X_14DAS	15928645	0.001[a]	212979611	0.000[b]	316624466	0.004[a]
0.125X_6DAS	15916495	0.002[a]	212975098	0.003[a]	316611985	0.000[a]
0.125X_0.5X	15914588	0.000[b]	212972959	0.002[a]	316613709	0.000[a]
n	199		227		766	
Bonf	0.000		0.000		0.000	
	GPAT8		***AOP3***		***AOP1***	
	SNP	P	SNP	P	SNP	P
Grand Mean	4185546	0.000[a]	41356523	0.000[b]	41344365	0.000[b]
Mean_14DAS	4184162	0.000[a]	41359617	0.000[b]	41356370	0.000[b]
Mean_6DAS	4177937	0.000[a]	41358745	0.000[b]	41343853	0.000[b]
Mean_0.5X	4184162	0.006[a]	41365317	0.000[b]	41356083	0.001[a]
0.25X_14DAS	4173542	0.001[a]	41359617	0.000[a]	41356370	0.001[a]
0.25X_6DAS	4167468	0.002[a]	41359167	0.000[a]	41346229	0.000[a]
0.25X_0.5X	4165676	0.000[a]	41356523	0.002[a]	41355061	0.001[a]
0.125X_14DAS	4179114	0.000[a]	41358745	0.000[b]	41341870	0.000[b]
0.125X_6DAS	4180957	0.000[a]	41358745	0.000[b]	41356395	0.000[b]
0.125X_0.5X	4179114	0.018[a]	41365317	0.000[a]	41344365	0.001[a]
n	457		236		248	
Bonf	0.000		0.000		0.000	

Table 3.18. Continued

	K23L20.6		*SPDS3*	
	SNP	P	SNP	P
Grand Mean	518038787	0.001[a]	521535920	0.000[a]
Mean_14DAS	518042751	0.002[a]	521535920	0.001[a]
Mean_6DAS	518043424	0.003[a]	521535920	0.000[b]
Mean_0.5X	518040172	0.001[a]	521535920	0.009[a]
0.25X_14DAS	518040457	0.010[a]	521534481	0.016[a]
0.25X_6DAS	518034068	0.001[a]	521535920	0.005[a]
0.25X_0.5X	518036335	0.000[b]	521533286	0.002[a]
0.125X_14DAS	518042714	0.001[a]	521535920	0.001[a]
0.125X_6DAS	518042556	0.020[a]	521535920	0.000[b]
0.125X_0.5X	518041166	0.008[a]	521531916	0.010[a]
n	683		332	
Bonf	0.000		0.000	

phenotype even though none of the SNPs within the gene showed significance in any of the phenotypes using the 1.6M SNPs dataset.

ALAAT1 encodes an enzyme that converts pyruvate and glutamate to alanine and 2-oxoglutarate (Liepman and Olsen 2003). However, the minor allele frequency (MAF) of the SNPs within the gene and the 20kb window surrounding the gene is very low (MAF < 5%). It is unclear if the significant effect seen using the 211K SNPs dataset was due to the low allelic frequency and population structure in that genomic region, or if the effect was true and *ALAAT1* played a role in glufosinate tolerance. The 1.6M SNPs dataset only contains SNPs with a moderate allele frequency (MAF > 5%), and none of the SNPs within the gene or within a 20kb window of the gene showed any significance except in the phenotype 0.125K_0.5X (Table 3.18). Although we potentially removed the SNP linked to the causative polymorphism, the linkage disequilibrium within the gene should cause other SNPs with a moderate MAF within the gene to have a significant association also, though it may be a smaller significance.

Since six of the eight genes showed significance, we tested the effect of each gene on glufosinate tolerance. We planted mutants for each of the putative genes and also of the *SHM* genes. We sprayed the mutants of the putative genes with 0.125X glufosinate. Only *ivd1-1* showed any significant difference between the control, Ler-0, even though none of the SNPs linked to *IVD1* showed any significance (Figure 3.6). We sprayed the *shm* mutants with 0.5X glufosinate and *shm1, shm3,* and *shm4* were significantly different between the control, Col-0 (Figure 3.7).

IVD1 is located in the mitochondria and is involved in leucine catabolism (Däschner et al. 1999, 2001). When *IVD1* was knocked out in seeds free amino acid

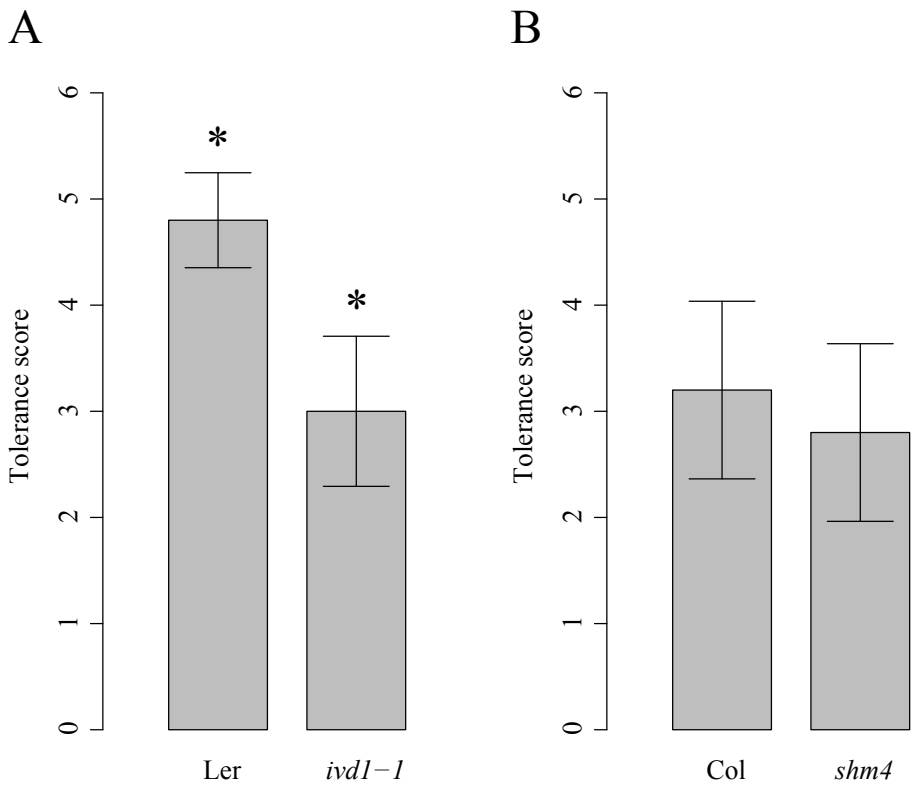

Figure 3.6. The average glufosinate damage score of *ivd1-1* and *shm4*. The y-axis is the damage scale: 1=healthy, no damage and 5=dead. The x-axis indicates the genotypes. Each genotype was sprayed with 0.125X glufosinate and scored 6 DAS. Statistical significance indicated by asterisk (p-value < 0.05). A) Comparing *ivd1-1* to wildtype, Ler-0. B) Comparing *shm4* to wildtype, Col-0.

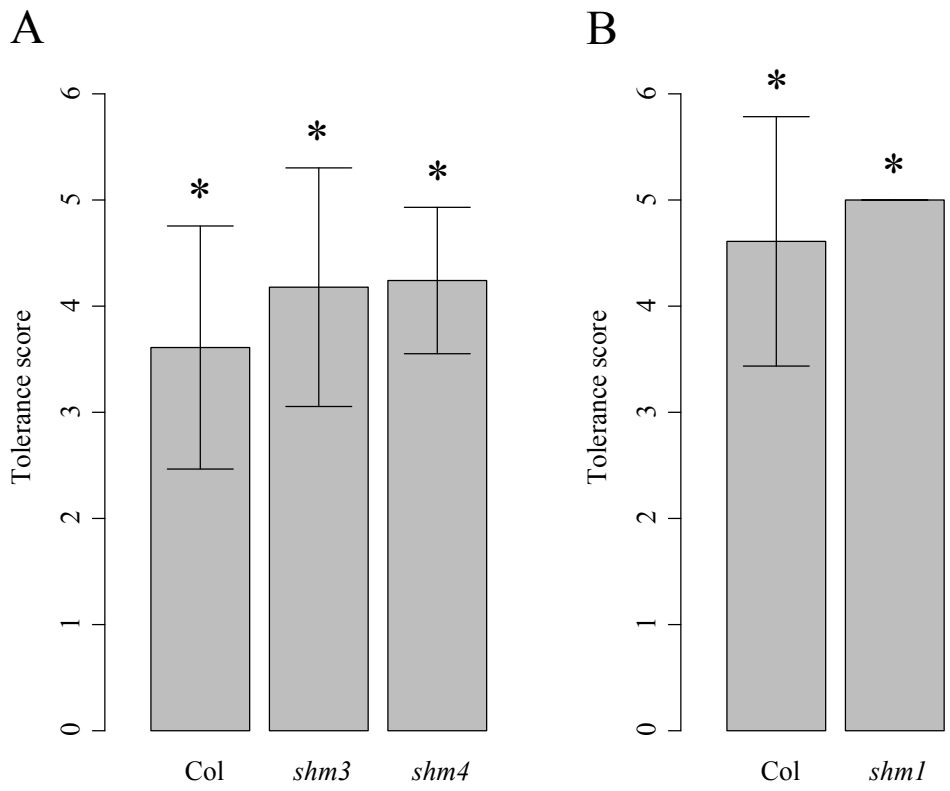

Figure 3.7. The average glufosinate damage score of *shm* mutants. The y-axis is the damage scale: 1=healthy, no damage and 5=dead. The x-axis indicates the genotypes The *shm* mutants were sprayed with 0.5X glufosinate and scored. Statistical significance is indicated by asterisks (p-value < 0.05). A) Damage scored 6 DAS. B) Damage scored 14 DAS.

levels were changed, and most amino acids concentrations increased dramatically (Gu et al. 2010). Also, the endogenous glucosinolate concentration was depleted, and a new glucosinolate was synthesized (Gu et al. 2010). We proposed that the free amino acids that accumulated in leaves in *ivd1-1* mutants were used during photorespiration and photosynthesis to compensate for the inhibition of GS2 similar to glufosinate rescue via amino acid application (Wendler et al. 1990). In addition, perturbation of glucosinolate biosynthesis perturbs amino acid biosynthesis (Chen et al. 2012). In *ivd1-1*, the glucosinolate biosynthesis was perturbed as was the amino acid concentrations. The effect of these perturbations on glufosinate tolerance are unknown and further study would be enlightening to understand the role of glucosinolate and amino acid biosynthesis in glufosinate tolerance.

3.2.2.2 Candidate genes and biological pathways affected by glufosinate

Next, we determined if the candidate genes and the genes involved in photorespiration and glucosinolate biosynthesis were affected by glufosinate. Abdeen et al. (2009) studied the effect glufosinate had on gene expression in wild type *A. thaliana* and genetically modified resistant *A. thaliana* (Abdeen and Miki 2009). We hypothesized that changes in gene expression could be an indication of putative genes contributing to glufosinate tolerance. We determined how many of the eight putative genes and the genes involved in photorespiration and glucosinolate synthesis changed gene expression upon glufosinate treatment.

Three of the eight putative genes changed expression in wild type *A. thaliana* after glufosinate treatment (Table 3.19). The genes that changed were *SPDS3*, *AOP1*, and

Table 3.19. The percentage of candidate genes that changed expression after glufosinate application (Abdeen and Miki 2009). The genes are separated into biological processes. The total column is the total number of genes involved in process. Genes involved in the glucosinolate biosynthesis and metabolic processes were found according to arabidopsis.org.

Genes	Total genes	Genes affected	Percentage
GWA candidate genes	8	3	37.5%
GLUTAMINE SYNTHETASE	6	5	83.33%
SERINE HYDROXYMETHYLTRANSFERASE	7	1	14.29%
Photorespiration	37	13	35.14%
Glucosinolate biosynthesis	169	88	52.07%
Glucosinolate Metabolic Processes	34	25	73.53%
Genome	36,310	3789	10.44%

GPAT8. Only *SHM1*, out of the seven *SHM* genes changed expression. Five of the six *GS* paralogs (only *GS1-5* did not change) changed expression. Also, 28 of the 38 genes involved in glucosinolate metabolic process were changed. Half of the 37 genes involved in photorespiration changed expression.

3.2.2.3 Candidate genes determined using MLMM

Interpretation of MLMM differs from EMMAX because the statistical model immediately eliminated the majority of SNPs that were found significant using the EMMAX model (Table 3.1). This resulted in very limited significant SNPs to analyze. To determine the relevance of the results, We determined how many of the eight putative genes, *GS* and *SHM* genes, and the photorespiration genes were found in the putative gene lists produced from the MLMM results. Only three of the eight putative genes were found in the candidate genes lists from the MLMM results, *AOP3*, *AOP1*, and *GPAT8*. None of the *GS* or *SHM* genes, and only one photorespiration gene was found in the candidate gene lists produced from the MLMM results.

Next, we asked how many of the MLMM putative genes were found in the lists of glufosinate-induced gene expression changes produced by the Abdeen et al. (2009) study. 43 candidate genes produced from the MLMM results were found to have changed gene expression after glufosinate tolerance. These 43 genes would be easy to follow-up and determine if inhibiting or suppressing these genes affected glufosinate tolerance.

3.2.3 Discussion

The natural variation of glufosinate damage was wide in *A. thaliana* (Figure 3.1). Using EMMAX and MLMM to calculate associations between genotype and phenotype

produced a large number of putative genes that could be contributing to glufosinate tolerance. The interpretation of the data was not easy, however. Using candidate gene lists comprising of putative genes, *GS* genes, *SHM* genes, and photorespiration genes, we found the lowest p-value within a 20kb window of those genes using the EMMAX results. Though these genes were good candidate genes as determined by the biological mechanism of glufosinate, none of these groups of genes showed major contribution to glufosinate tolerance in *A. thaliana*.

It was difficult to use the MLMM results for the same test because the associations were calculated with SNPs added to the model as cofactors. This meant that the association of each SNP was not independent of each other, and interpreting *p*-values was not as clear as with EMMAX results.

I used a second method used to test the plausibility of the candidate genes involved in biological pathways using data of gene expression changes after a glufosinate treatment (Abdeen and Miki 2009). Five of the six *GS* genes, half of the photorespiration, and a majority of the glucosinolate metabolic process genes changed expression after glufosinate treatment. Only three of the eight putative genes were found on the list. These results would indicate that nitrogen assimilation, photorespiration, and glucosinolate biosynthesis is greatly altered after a glufosinate treatment. Changes in these biological processes could potentially lead to tolerance if these changes allow the plant to continue photosynthesis.

Determining the significant SNPs that are true positives, and which genes linked to those SNPs are the true genes is difficult. Neither MLMM nor EMMAX gave obvious genes that contributed to the variation of glufosinate tolerance in *A. thaliana*. Using

significant cutoffs determines the ratio of false verses true positives are analyzed. For example, even though SNPs linked to *IVD1* and the *SHM* genes did not have significantly low p-values, and they were not found in the lists of genes that changed expression after glufosinate treatment, the mutants of these genes were the only ones that had a significantly different response to glufosinate than the controls. After the spraying *shm* mutants and mutants of the putative genes, the *ivd1-1* mutant showed significant decrease in glufosinate damage compared to the Ler-0, and *shm3*, *shm4,* and *shm1* mutants showed significant increase in glufosinate damage (Figures 3.6 & 3.7). These results suggested that these genes were playing a role in glufosinate tolerance in *A. thaliana* even though the results from the EMMAX and MLMM results did not suggest it. Therefore, it was difficult to interpret the results, and deciding which candidate genes to follow-up on is a difficult matter.

3.3 Hybrid incompatibility

Hybrid incompatibility was phenotyped based on the rate of seed lethality of hybrid seeds between *A. thaliana* and *A. arenosa*. These two species diverged from each other ~5 million years ago (Koch et al. 2000, Kuittinen and Aguadé 2000). Even though the hybridization barrier is high between these two species, a natural hybrid, *A. suecica*, does exist (Kamm et al. 1995, O'Kane et al. 1996, Josefsson et al. 2006). The genetic differences between the three species has been highly studied to understand the genetic consequences of interspecific hybridizations and allopolyploidization (Comai et al. 2000, Bushell et al. 2003, Josefsson et al. 2006, Chang et al. 2010, Burkart-Waco et al. 2012, 2013).

To further understand the genetic mechanisms underlying the hybridization barriers between these two species, we crossed 440 *A. thaliana* accessions with *A. arenosa*. The seed abortion rate varied among *A. thaliana* accessions, and we hypothesized that the natural variance could be mapped using GWA to find candidate genes that contributed to post-hybridization barrier (Dilkes et al. 2008, Burkart-Waco et al. 2012).

3.3.1 Methods

Using fine-tipped forceps, all open flowers were removed from a mature branch. Unopened flowers were emasculated by removing the sepals, petals, and stamen, leaving only the pistil. Stamens from *A. arenosa* were picked and dapped or wiped onto the exposed *A. thaliana* pistil with the pollen. The exposed pistil was left opened, but separated from all other flowers to eliminate pollen-contamination. The pistils were repollinated the next morning. Repollination greatly increased the success of the crosses. All pollinations were done in the morning and early afternoon, as this was the best time for emasculations and pollinations. Three different flower clusters were pollinated in order to get replicates and to get a true count of seed phenotypes.

After seed maturation (18-21 days after pollination) the seeds were collected. The seeds were then cleaned from the siliques and were categorized into four different categories: plump, shriveled, green, and viviparous. The plump seeds were mostly-normal looking seeds, some of them differing in size. The shriveled seeds had no endosperm and no or very small embryo, but the seed coat had developed. The green seeds looked like a normal seed, except the seed coat never finished desiccation and

maintained a green coloration. The viviparous seeds were normally developed, but they germinated early.

These four phenotypes captured different developmental stages: abortion before embryogenesis (shriveled), no seed maturation (green), no dormancy (viviparous), and complete seed maturation (plump). In addition to these four phenotypes different ratios of these categories were calculated to capture complete developmental stages (Table 3.20) (Burkart-Waco et al. 2012). The set of phenotypes was run through the pipeline and putative gene lists are available.

3.3.2 Results

The EMMAX results using the 211K and 1.6M SNPs datasets differed from each other (Figures 3.8 & 3.9). The most significant SNP using the 211K SNPs dataset for %P, Chr1:14704182, was not included in the 1.6M SNPs dataset, but the SNPs around that region did not show any significance in the results. The 1.6M SNPs dataset results showed a lot more significant regions than the 211K SNPs dataset results.

The EMMAX results differed from the MLMM results. The two model selection criteria of MLMM produced the same model for %P (Figures 3.10 & 3.11). MLMM produced a model that contained five significant SNPs (Table 3.20), and one of those SNPs was Chr1:14704182.

The EMMAX results gave a large number of significant SNPs, and using MLMM greatly reduced the number of significant SNPs (Table 3.20). One method to select genes of interest for further study is to compare the GWA results to genes discovered in other studies as having a potential role in normal and hybrid seed development

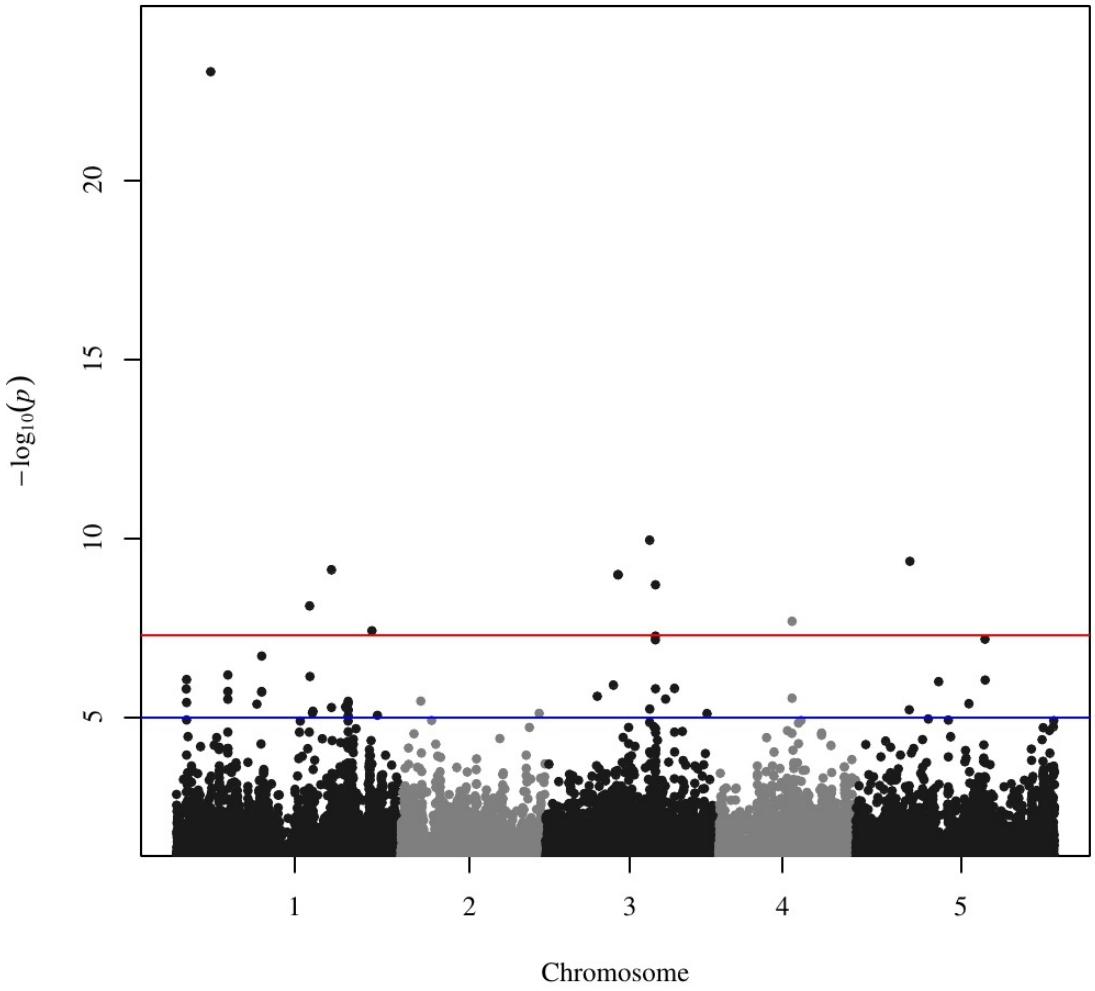

Figure 3.8. Manhattan plot displaying the *p*-values of 211K SNPs calculated using EMMAX for the phenotype %P. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The blue line indicates suggested significant cutoff ($\alpha \leq 1x10^{-5}$) and the red line indicates a genomewide significant cutoff ($\alpha \leq 5x10^{-8}$).

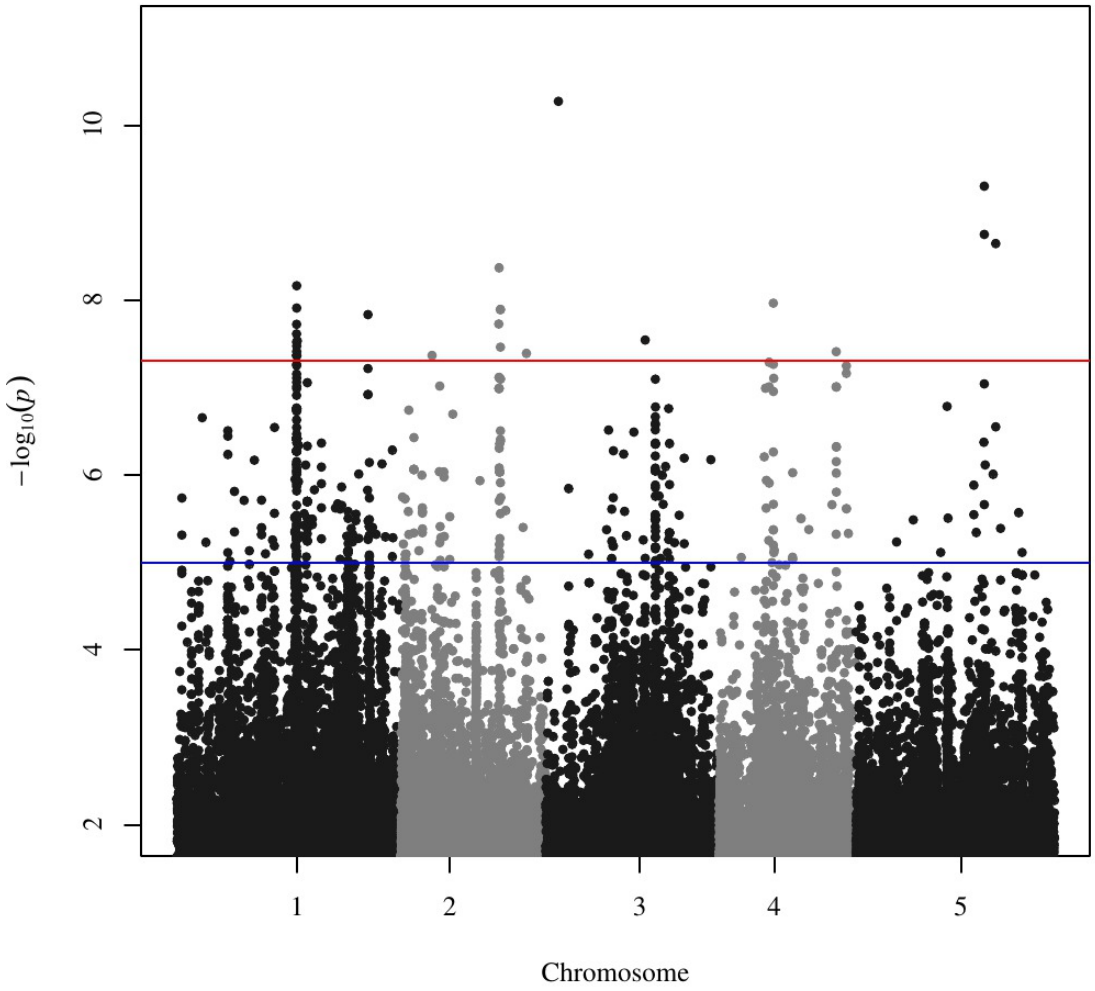

Figure 3.9. Manhattan plot displaying the *p*-values of 1.6M SNPs calculated using EMMAX for the phenotype %P. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The blue line indicates suggested significant cutoff ($\alpha \leq 1x10^{-5}$) and the red line indicates a genomewide significant cutoff ($\alpha \leq 5x10^{-8}$).

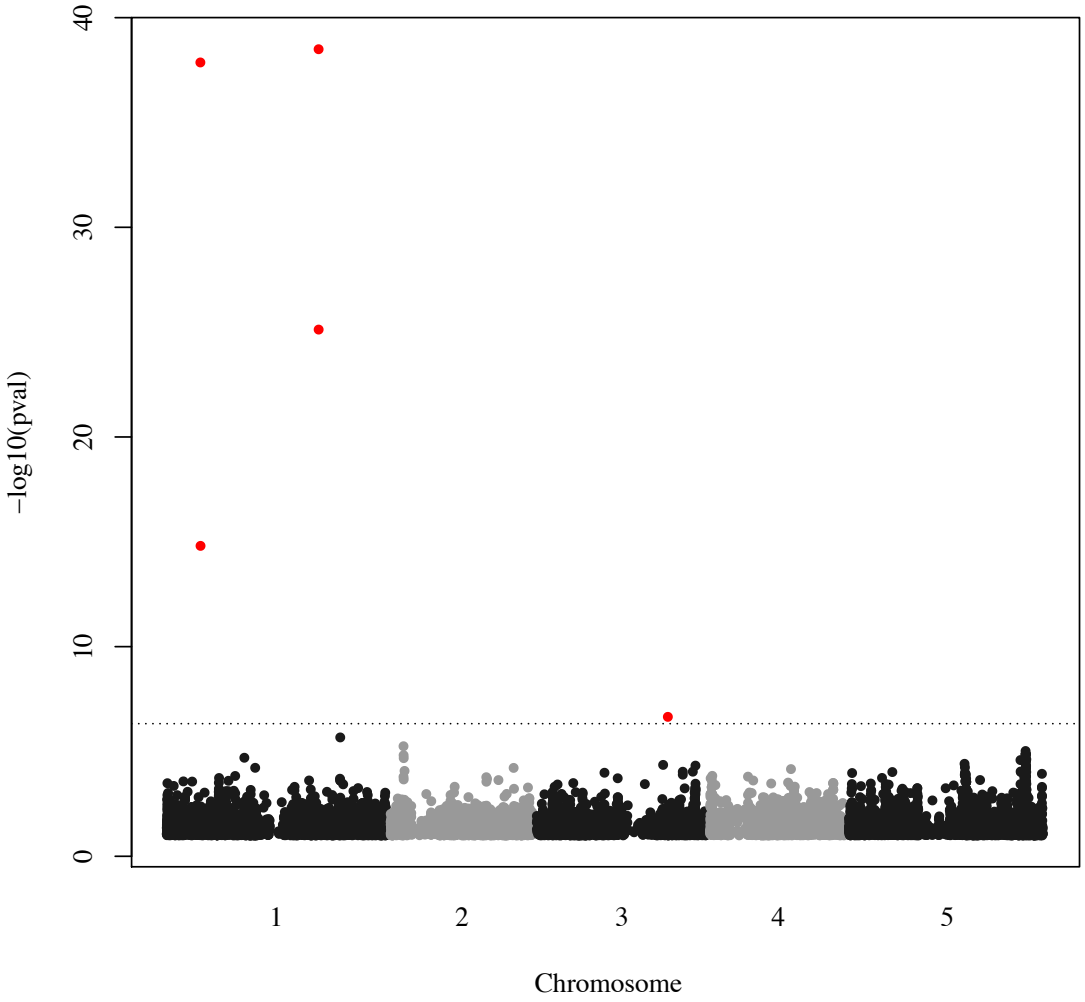

Figure 3.10. Manhattan plot displaying the *p*-values of 211K SNPs calculated using MLMM for the phenotype %P. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The red highlighted SNPs are the significant SNPs as determined by the model selection criterion BIC.

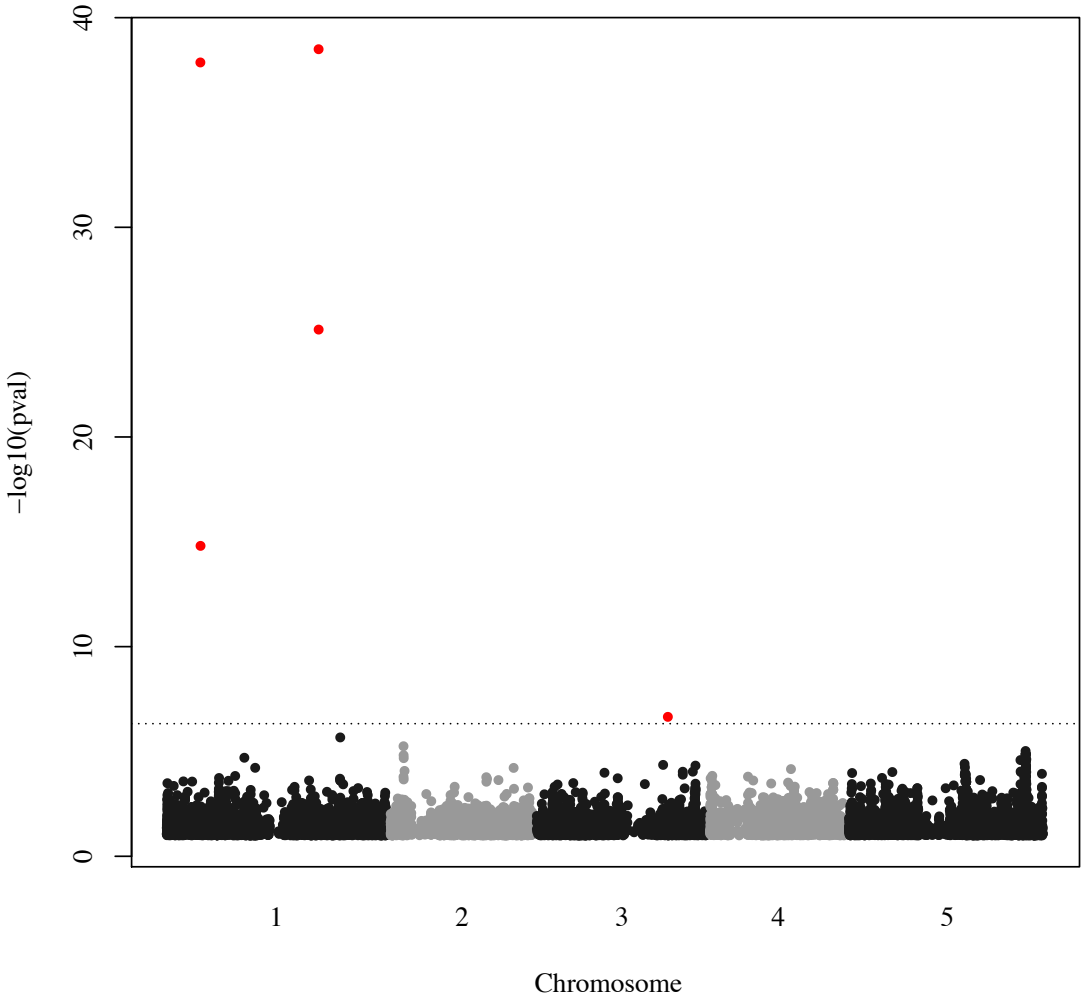

Figure 3.11. Manhattan plot displaying the *p*-values of 211K SNPs calculated using MLMM for the phenotype %P. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The red highlighted SNPs are the significant SNPs as determined by the model selection criterion Bonferroni.

Table 3.20. The number of significant SNPs from the two statistical models: EMMAX and MLMM for phenotypes of hybridization incompatibility using *A. thaliana* and *A. arenosa* as parents. Significance for the EMMAX results was defined as α (<= $1x10^{-5}$).

Phenotype	EMMAX 211K	EMMAX 1.6M	MLMM BIC	MLMM BONF
%G/PVG	2	4	0	0
%Green	5	48	1	1
%P	49	465	5	5
%P/PVG	0	5	0	0
%P/PV	2	50	0	0
%PVG	13	48	2	3
%PV	22	291	4	4
%SG	69	626	4	6
%SGV	98	785	8	7
%Shrivel	13	132	3	3
%V/PVG	11	105	0	0
%VG	6	48	1	1
%Vivi	16	109	0	1

(Atwell et al. 2010). This approach was used for selecting genes of interest contributing to hybrid incompatibilities.

A candidate gene list was created based on literature consisting of 21 genes (Table 3.21). These genes included proanthocyanidin biosynthetic genes, DNA methylation regulatory genes, seed development regulatory development, and chromatin assembly. Of these 21 genes, only two genes were found in the candidate gene list results from the 211K SNPs EMMAX results and six were found in the candidate gene list results from the 1.6M SNPs EMMAX results (Tables 3.22 & 3.23). None of the genes were found in the MLMM analyses.

Next, three different studies were used to determine candidate genes, each one looking at different genetic aspects upon pollen fertilization to determine which putative genes generated through GWA might be of interest (Nodine and Bartel 2012, Burkart-Waco et al. 2013, Schatlowski et al. 2014).

The first study measured the change of gene expression in *A. thaliana* x *A. arenosa* hybrid seeds by comparing the gene expression of the hybrid seeds to both parents (Burkart-Waco et al. 2013). Burkart-Waco et al. (2013) categorized gene expression change into four categories: Changed gene expression compared to *A. arenosa* only, changed gene expression compared to total *A. arenosa*, changed gene expression compared to *A. thaliana*, and changed gene expression compared to both parents (Burkart-Waco et al. 2013). The results given by EMMAX using the 211K SNPs dataset had ~10% of the genes for each respective category, and the EMMAX results using the 1.6M SNPs dataset contained more than 30% of the genes that changed expression in all four different categories (Table 3.24). MLMM results, on the other hand, contained none

Table 3.21. List of 21 candidate genes selected by literature.

Locus	**Gene Name**	**Function**
AT1G65470	*FAS1*	Chromatin assembly
AT5G49160	*MET1*	Methylation
AT1G02580	*MEDEA*	PRC2
AT2G35670	*FIS2*	PRC2
AT3G12280	*RBR*	PRC2
AT5G58230	*MSI1*	PRC2
AT1G17260	*AHA10*	Proanthocyanidin biosynthesis
AT1G61720	*BAN*	Proanthocyanidin biosynthesis
AT2G37260	*TTG2*	Proanthocyanidin biosynthesis
AT3G51240	*TT6*	Proanthocyanidin biosynthesis
AT3G55120	*TT5*	Proanthocyanidin biosynthesis
AT3G59030	*TT12*	Proanthocyanidin biosynthesis
AT4G09820	*TT8*	Proanthocyanidin biosynthesis
AT4G22880	*TT18/TDS4*	Proanthocyanidin biosynthesis
AT5G07990	*TT7*	Proanthocyanidin biosynthesis
AT5G13930	*TT4*	Proanthocyanidin biosynthesis
AT5G24520	*TTG1*	Proanthocyanidin biosynthesis
AT5G35550	*TT2*	Proanthocyanidin biosynthesis
AT5G42800	*TT3*	Proanthocyanidin biosynthesis
AT5G48100	*TT10/LAC15*	Proanthocyanidin biosynthesis

Table 3.22. The number of candidate genes found in putative gene lists for the seed lethality phenotypes produced by EMMAX and MLMM. Each analysis is represented by two columns: All = All genes that are linked to the significant SNPs, the hit gene and the 10 genes that are up- and downstream of the significant SNP; Hit= Only looking at the "Hit" genes, the genes that contained the SNP or was the closest to the SNP if the SNP fell out of a gene.

	Total	**EMMAX 211K**		**EMMAX 1.6M**		**MLMM**	
		All	Hit	All	Hit	All	Hit
Candidate Genes	21	2	0	6	2	0	0

Table 3.23. List of the candidate genes found in the putative gene lists for the hybrid incompatibility phenotypes from the EMMAX and MLMM results. The phenotypes, in which the gene was found, are indicated by the analysis. The candidate genes that were "Hit" genes are indicated by bolded print.

Gene	**Name**	**EMMAX 211K**	**EMMAX 1.6M**	**MLMM**
AT1G17260	*AHA10*	%SG; %SGV	%SG; %SGV	-
AT1G61720	*BAN*	%SGV	-	-
AT1G65470	*FAS1*	-	**%P**; %P/PV; %SG; **%SGV**; %V	-
AT2G35670	*FIS2*	-	%P/PVG; %V/PVG	-
AT4G09820	*TT8*	-	%P; %PV; %SGV; %SG	-
AT5G42800	*TT3*	-	%P; **%SGV**	-
AT5G48100	*TT10*	-	%P/PV	-

Table 3.24. The number of genes that have changed gene expression after interspecies hybridization found in putative gene lists for the seed lethality phenotypes produced by EMMAX and MLMM (Burkart-Waco et al. 2013). The total number of genes that changed expression after hybridization is indicated by the column "Total." Each analysis is represented by two columns: All = All genes that are linked to the significant SNPs, the hit gene and the 10 genes that are up- and downstream of the significant SNP; Hit= Only looking at the "Hit" genes, the genes that contained the SNP or was the closest to the SNP if the SNP fell out of a gene.

	Total	**EMMAX 211K**		**EMMAX 1.6M**		**MLMM**	
		All	Hit	All	Hit	All	Hit
Hybrid vs AaOnly	3298	396	56	1202	216	18	2
Hybrid vs AaTotal	3592	429	60	1320	231	19	2
Hybrid vs Athaliana	214	29	0	80	12	0	0
Hybrid vs BothParents	293	33	5	118	87	1	0

or a very small percentage of genes that changed expression in hybrid seeds compared to the parental gene expression (Table 3.24). Any of those genes found by Burkart-Waco et al. (2013) are putative genes that could be studied to determine role in hybridization barriers. However, there are far too many still to study, so to decrease candidate genes, We looked at how many of the candidate genes found in the GWA results either contained the significant SNP or was the closest gene to the SNP, as indicated by the term "Hit." This greatly decreased the number of candidate genes and suggested that change in gene expression was not a great indicator of genes playing a function in the seed hybrid incompatibilities (Table 3.24).

Burkart-Waco et al. (2013) also comprised their own list of candidate genes based on their gene expression analysis. They determined that 28 genes involved in defense response or transcription regulation were good candidates for contributing to the hybridization barrier (Burkart-Waco et al. 2013). Of these 28 candidate genes, three were found in the candidate genes lists of the phenotypes using 211K SNPs dataset and 13 of them were found in the 1.6M SNPs dataset using the EMMAX method (Tables 3.25 & 3.26). None of the candidate genes was found using MLMM (Table 3.25).

Secondly, we asked the question if any of the putative genes determined via EMAMX or MLMM were found to be differentially methylated in interploidy hybrid seeds (Schatlowski et al. 2014). Genomic methylation plays a vital role in gene expression and has been shown to change upon hybridizations (Adams et al. 2000, Comai et al. 2000, Josefsson et al. 2006, Chen et al. 2008, Chang et al. 2010, Xiang et al. 2011, Schatlowski et al. 2014). Schatlowski et al. (2014) comprised lists of genes that either lost or gained methylation depending on different pollen parents. 43% of the genes that

Table 3.25. The number of candidate genes contributing to hybrid incompatibility determined by Burkart-Waco et al. (2013) found in putative gene lists for the seed lethality phenotypes produced by EMMAX and MLMM. The total number of genes that changed expression after hybridization is indicated by the column "Total." Each analysis is represented by two columns: All = All genes that are linked to the significant SNPs, the hit gene and the 10 genes that are up- and downstream of the significant SNP; Hit= Only looking at the "Hit" genes, the genes that contained the SNP or was the closest to the SNP if the SNP fell out of a gene.

	Total	**EMMAX 211K**		**EMMAX 1.6M**		**MLMM**	
		All	Hit	All	Hit	All	Hit
Candidate Genes	28	3	0	13	3	0	0

Table 3.26. List of the candidate genes determined by Burkart-Waco et al. (2013) found in the putative gene lists produced by EMMAX and MLMM. A column represents each analysis, and the phenotype in which the gene was found was listed by analysis. The candidate genes that were "Hit" genes are indicated by bolded print.

Gene	Name	EMMAX 211K	EMMAX 1.6M	MLMM
AT1G15800		%V		-
AT4G16845	*VRN2*	%P/PV	%P; **%G**; %P/PVG; %P/PV; **%PVG**; %SG; **%VG**	-
AT5G33340	*CDR1*	%P; %SGV; %SG	%P; %PV; %SGV; %SG	-
AT1G22590	*AGL87*	-	%P; %PVG; %PV; %SGV; %SG; **%S**; %VPVG; **%V**	-
AT1G34650	*HDG10*	-	%P; %VG; %V	-
AT1G64280	*NPR1*	-	%P; %PV; %SG	-
AT1G65330	*PHE1*	-	%P; %SGV	-
AT1G74710	*EDS16*	-	%SGV	-
AT2G35670	*FIS2*	-	%P/PVG; %P/PV; %V/PVG	-
AT2G40210	*AGL48*	-	%P/PVG	-
AT3G44630		-	%P/PVG	-
AT5G60440	*AGL62*	-	%P; %SGV; %SG	-
AT5G62165	*AGL42*	-	**%P**; **%PVG**; **%PV**; **%SGV**; **%SG**; **%S**	-
AT5G65050	*AGL31*	-	%SGV; %V/PVG	-

gained methylation using a hypomethylated (*met1*) pollen parent were found in the list of putative genes of the 1.6M SNPs dataset (Table 3.27). 58% of the genes that lost methylation using a tetraploid (*osd1*) pollen parent were found in the putative gene list of the 1.6M SNPs dataset. 45% of the genes that gained methylation using the hypomethylated tetraploid (*osd1met1*) pollen parent were found in the putative gene list of the 1.6M SNPs dataset. Only 13-19% of the genes that had changed methylation patterns were found in the putative genes of the 211K SNPs dataset. Once again, a very small percentage of these genes were found in the putative gene lists of the MLMM models.

Lastly, we compared the putative gene lists created using EMMAX and MLMM to a list of genes that are differentially expressed in the early zygotic stages (Nodine and Bartel 2012). Nodine and Bartel (2012) found that the maternal and paternal genomes contributed equally in early zygotic development, and only 122 genes were differentially expressed during the 1/2-cell stage, the 8-cell stage or 32-cell stage. 40% of these differentially expressed genes were found in the EMMAX results using the 1.6M SNPs dataset and only 14% using the 211K SNPs dataset (Table 3.28). Only one gene was found using MLMM.

3.3.3 Discussion

As one method to interpret the GWA results, we compared the putative gene lists created through the pipeline to a list of candidate genes based on literature as showing to affect hybridization barriers. These genes consist of proanthocyanidin biosynthesis genes,

Table 3.27. The number of genes with changed methylation in interploidy hybridizations found in putative gene lists for the seed lethality phenotypes produced by EMMAX and MLMM (Schatlowski et al. 2014). The total number of genes that changed expression after hybridization is indicated by the column "Total." Each analysis is represented by two columns: All = All genes that are linked to the significant SNPs, the hit gene and the 10 genes that are up- and downstream of the significant SNP; Hit= Only looking at the "Hit" genes, the genes that contained the SNP or was the closest to the SNP if the SNP fell out of a gene.

	Total	**EMMAX 211K**		**EMMAX 1.6M**		**MLMM**	
		All	Hit	All	Hit	All	Hit
CHG_gain_*met1*	281	38	8	120	21	2	1
CHH_loss_*osd1*	1847	355	53	1072	252	11	1
CHG_gain_*osd1met1*	427	60	8	192	44	5	0

Table 3.28. The number of genes that are differentially expressed in early zygotic development found in the putative gene lists for the seed lethality phenotypes produced by EMMAX and MLMM (Nodine and Bartel 2012). The total number of genes that changed expression after hybridization is indicated by the column "Total." Each analysis is represented by two columns: All = All genes that are linked to the significant SNPs, the hit gene and the 10 genes that are up- and downstream of the significant SNP; Hit= Only looking at the "Hit" genes, the genes that contained the SNP or was the closest to the SNP if the SNP fell out of a gene.

	Total	**EMMAX 211K**		**EMMAX 1.6M**		**MLMM**	
		All	Hit	All	Hit	All	Hit
Imprinted Genes	122	17	3	49	10	1	0

chromatin rearranging genes, and seed development regulatory genes (Table 3.21). These were not heavily represented in the GWA results.

I also compared the putative gene lists to candidate gene lists created based on genomic changes during normal, interspecific, and interploidy seed development. Three different aspects of seed development were analyzed: changed gene expression due to hybridization, changed methylation patterns due to hybridization, and parentally differentiated gene expression during early zygotic development (Nodine and Bartel 2012, Burkart-Waco et al. 2013, Schatlowski et al. 2014). The GWA results did not contain many genes that experienced gene expression changes after hybridization (Table 3.24). This would indicate that the variation that contributed to hybrid incompatibilities in *A. thaliana* was not linked to changes in gene expression. However, genes that were differentially expressed during the early zygotic stages and genes that undergo changes in methylation patterns were highly represented in the EMMAX results using the 1.6M SNPs dataset (Tables 3.27 & 3.28).

Neither the EMMAX results using the 211K SNPs dataset or the MLMM results produced any obvious putative genes based on the comparisons between the GWA putative gene lists and the candidate gene lists. These two analyses produced limited results compared to the EMMAX results using the 1.6M SNPs dataset. Comparing the results produced by EMMAX using the two SNP datasets, 211K SNPs did not produce as many results as the 1.6M SNPs dataset, and the putative genes differed between the two SNP datasets. This could indicate that one set of SNPs is more correct than the other.

The kinship was calculated for each SNP dataset and if the 211K SNPs did not represent the population structure correctly than the EMMAX results would be biased.

The same is true for the 1.6M SNPs. If, by removing all SNPs with MAF ≤ 5%, changed the kinship file so that it no longer represented the true kinship between accessions then the results from the 1.6M SNPs dataset would be biased (Table 3.20).

Comparing MLMM to EMMAX, MLMM produced a very small number compared to EMMAX. This could indicate that either MLMM found new/additional genes that are contributing to the hybridization barriers, and they are the potentially interesting putative genes, or that MLMM was not a good model choice for these phenotypes.

3.4 Seed size

The seed sizes of the 440 accessions of *A. thaliana* were measured with the purpose of studying the correlation between size and the severity of the hybridization barrier. It was hypothesized that seed size contributed to hybridization barrier because of the nature of resource allocation from sporophyte to developing zygote (Haig and Westoby 1991). We hypothesized that if seed size contributed to the hybridization barriers than similar SNPs and genes should be significant in both GWA analyses of hybrid incompatibilities and seed size. These results should indicate which genes are playing a role in both seed size and the hybridization barrier and should indicate more clearly how seed size is contributing to the hybridization barrier. Seed size was described as different measuring methods (Table 3.29). Also, to continue testing the hypothesis that the proanthocyanidin biosynthesis pathway may contribute to hybridization barriers, the difference of coloration was calculated by measuring the amount of red, green, and blue in the seeds.

3.4.1 Methods

Approximately 100 Seeds from each accession were spread out onto a scanner, one accession at a time and a picture of the seeds were scanned. Three replicate scans of three different groups of seeds for each accession were taken so that the size variation within each accession was calculated. Seed size was calculated computationally by counting the number of pixels that comprised the seed.

The BLUPs for each accession for each phenotype was calculated in R (R Core Team 2013). The BLUPS were used to represent each size phenotype that was run through the pipeline.

3.4.2 Results

Using the manhattan plots to visualize the GWA results showed that the results differed significantly depending on the statistical method used and the number of SNPs used (Figures 3.12-3.15). Once again, EMMAX, using the 1.6M SNPs dataset, gave more significant SNPs than either the results from the 211K SNPs dataset using EMMAX or MLMM (Table 3.29). Even though MLMM results contained many SNPs that had a *p*-value $< 1x10^{-5}$ only one SNP was found significant using the Bonferroni model selection criterion (Figures 3.14 & 3.15; Table 3.29).

I hypothesized that seed size and seed lethality phenotypes should share significant SNPs in the EMMAX analyses if the two traits were correlated. To test this, we compared the *p*-values of the significant SNPs from each seed lethality phenotype to the *p*-values calculated for each seed size phenotype. We also did the reverse, comparing the *p*-values of the significant SNPs for seed size to the *p*-values calculated for each seed

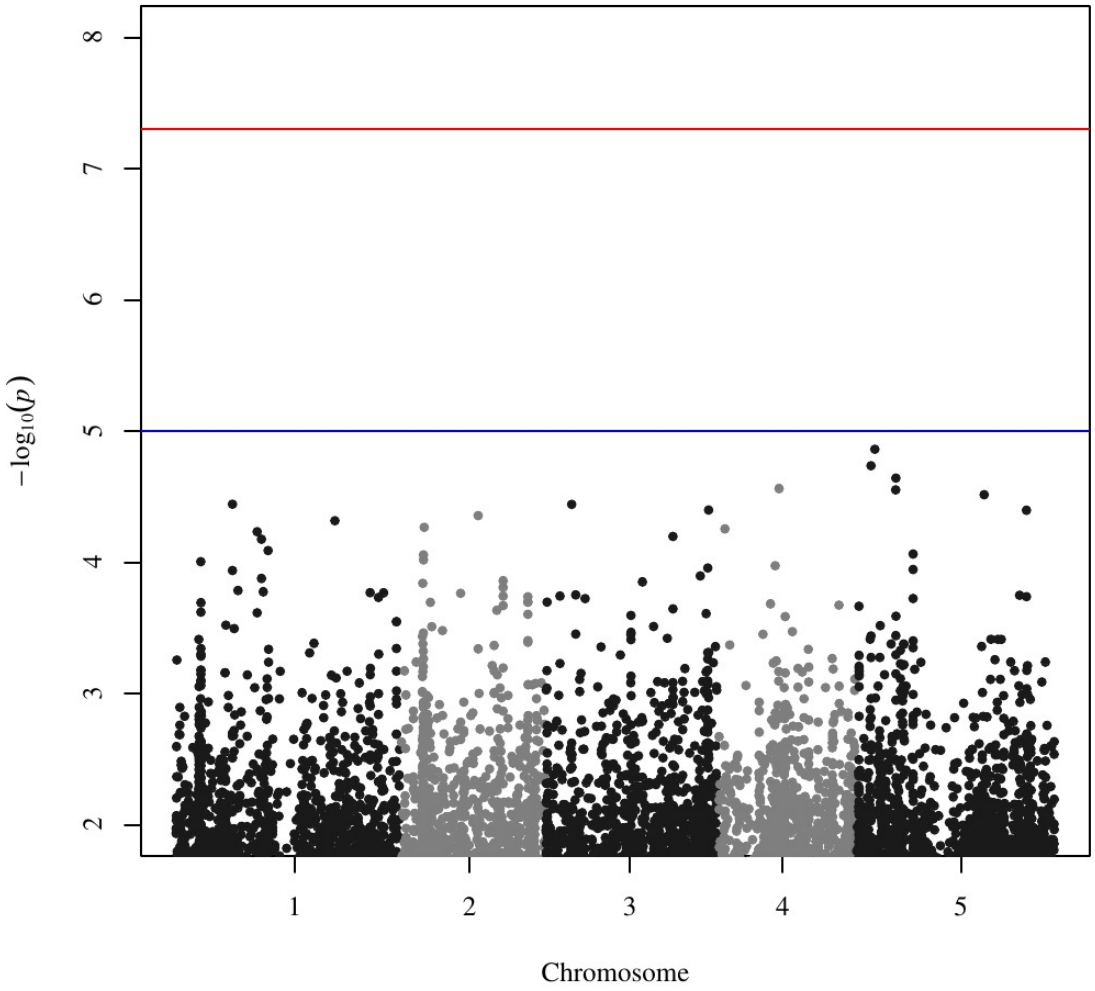

Figure 3.12. Manhattan plot displaying the *p*-values of 211K SNPs calculated using EMMAX for the phenotype Average Area. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The blue line indicates suggested significant cutoff ($\alpha \leq 1x10^{-5}$) and the red line indicates a genomewide significant cutoff ($\alpha \leq 5x10^{-8}$).

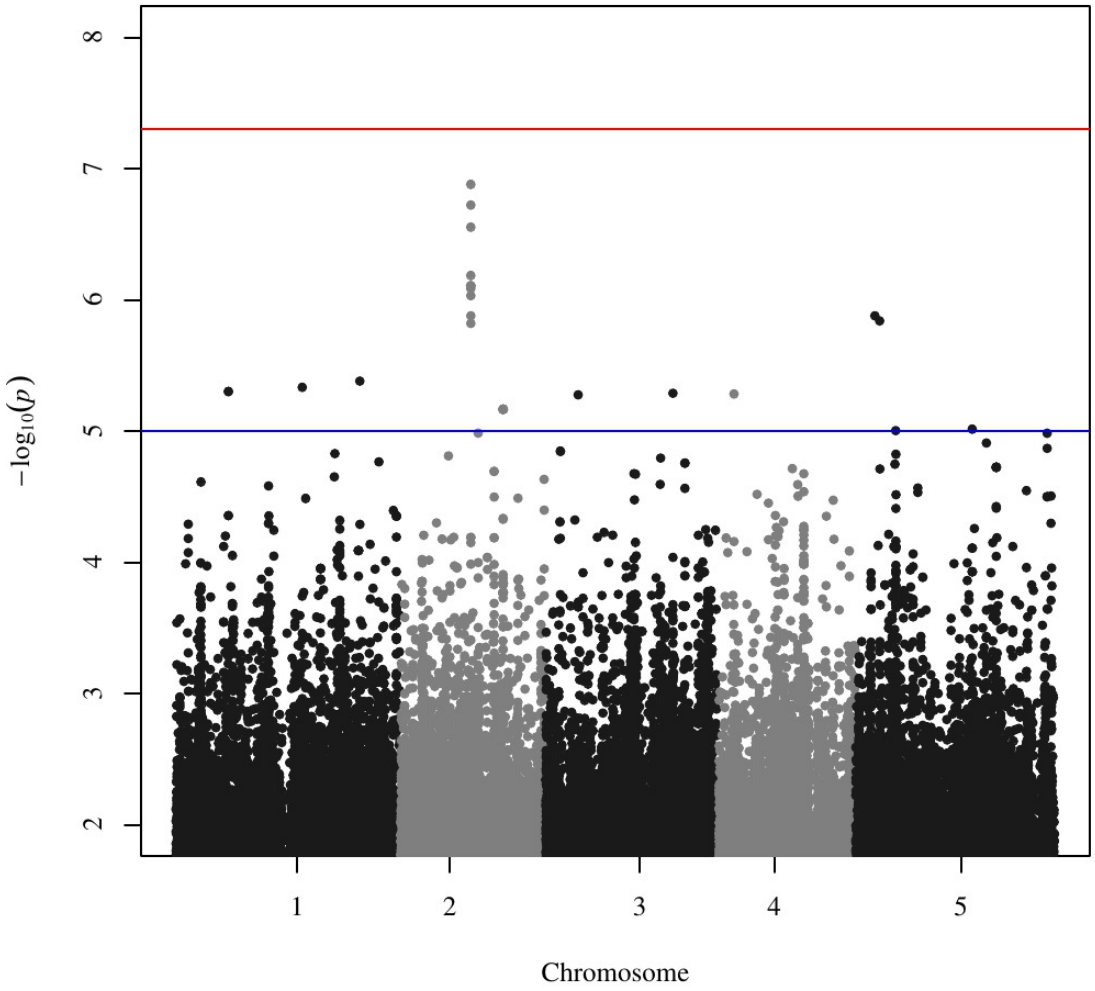

Figure 3.13. Manhattan plot displaying the *p*-values of 1.6M SNPs calculated using EMMAX for the phenotype Average Area. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The blue line indicates suggested significant cutoff ($\alpha \leq 1x10^{-5}$) and the red line indicates a genomewide significant cutoff ($\alpha \leq 5x10^{-8}$).

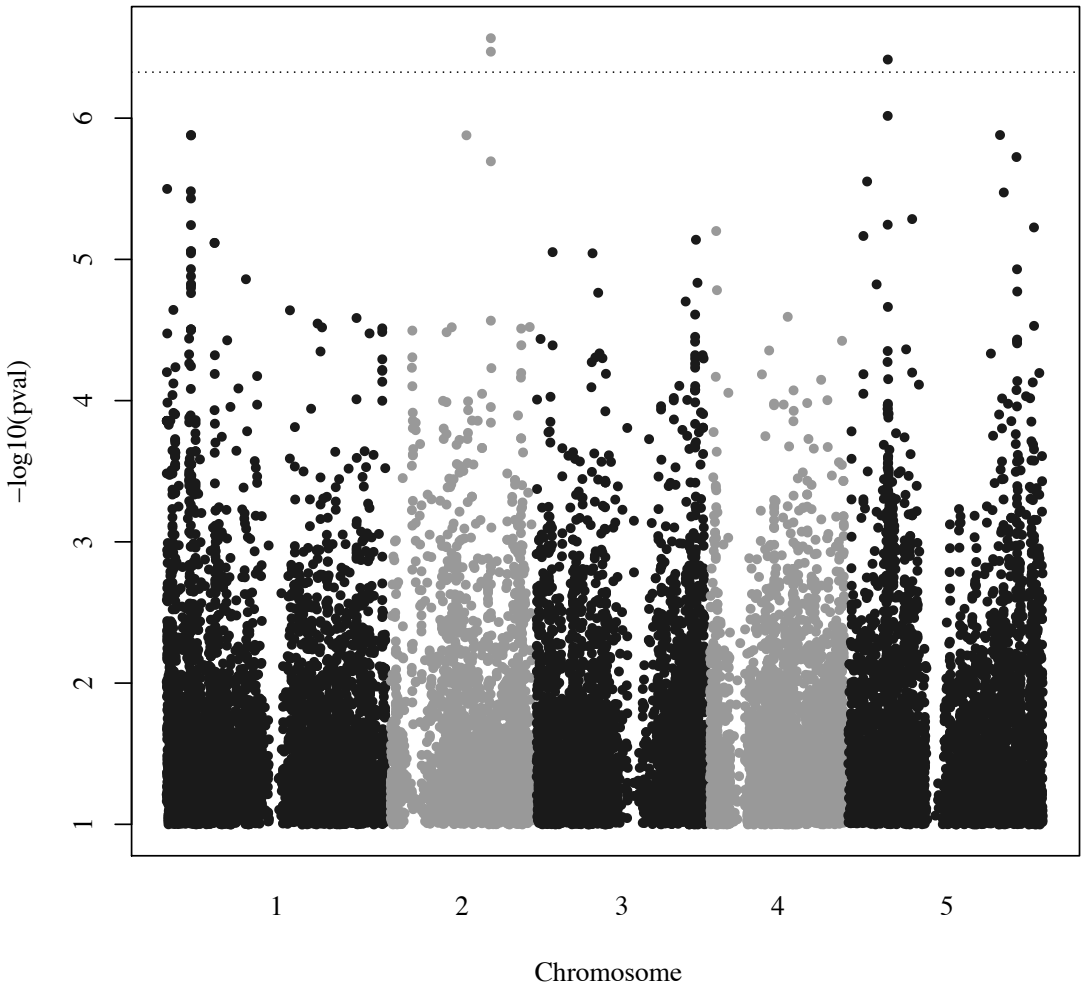

Figure 3.14. Manhattan plot displaying the *p*-values of 211K SNPs calculated using MLMM for the phenotype Average Area. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The red highlighted SNPs are the significant SNPs as determined by the model selection criterion BIC.

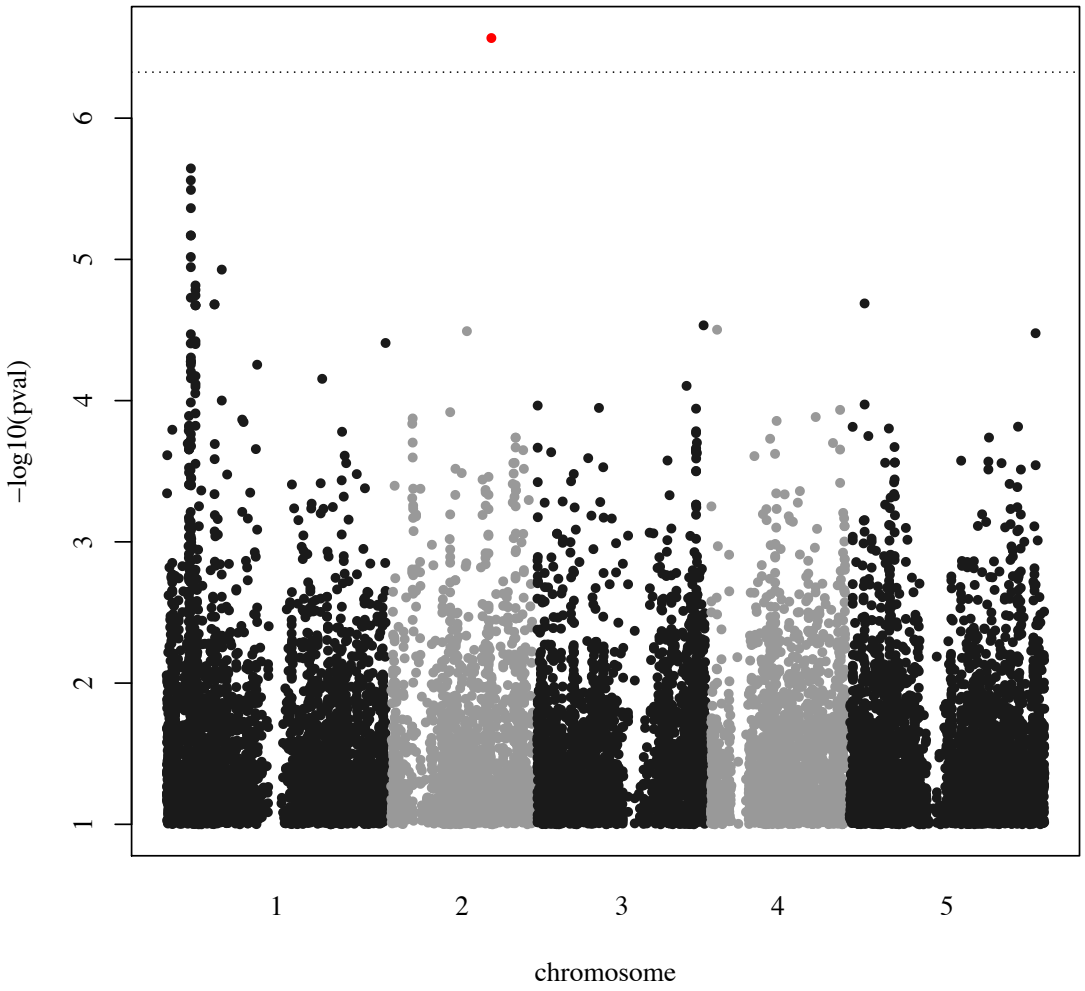

Figure 3.15. Manhattan plot displaying the *p*-values of 211K SNPs calculated using MLMM for the phenotype Average Area. The y-axis is the *p*-value transformed to the negative log. The x-axis annotates the five chromosomes of *A. thaliana*, coloration indicates the start and end of the chromosomes. The red highlighted SNPs are the significant SNPs as determined by the model selection criterion Bonferroni.

Table 3.29. The number of significant SNPs from the two statistical models: EMMAX and MLMM for phenotypes of seed size. Significance for the EMMAX results was defined as ($\alpha <= 1x10^{-5}$).

Phenotype	EMMAX 211K	EMMAX 1.6M	MLMM BIC	MLMM BONF
Average Area	0	30	1	1
Average Circumference	2	10	4	4
Average Perimeter	2	30	0	1
Average Area/Perimeter	0	26	0	1
Average Radius Average	0	38	0	1
Average Radius max.	0	29	1	2
Average Radius min.	2	30	0	1
Average MIB	34	124	1	3
Average MIG	48	183	3	3
Average MIR	31	115	3	3
Average STDIB	15	76	0	4
Average STDIG	23	125	3	3
Average STDIR	38	156	2	2
MIG/MIB	14	86	1	1
MIR/MIB	20	49	1	2
MIR/MIG	10	52	1	1

lethality phenotype. Comparing the two sets of SNP *p*-values, seed size and seed lethality do not share common significant SNPs (Figures 3.16-3.19).

3.4.3 Discussion

Although seed size was hypothesized to contribute to the hybridization barrier, there was no correlation between the results produced by EMMAX for seed lethality and seed size. The significant SNPs from the seed lethality analysis did not share significance with seed size, and the significant SNPs from the seed size analysis were not significant in the seed lethality phenotypes. Therefore, there were no obvious common genes that linked seed size with the hybridization barrier.

3.5 Secondary Metabolites

Secondary metabolites are numerous and little is known about the biosynthesis and the function of most metabolites (Dixon and Strack 2003, D'Auria and Gershenzon 2005, Yonekura-Sakakibara and Saito 2009). To facilitate the metabolic research, stem and leaf tissues of 440 accessions of *A. thaliana* were harvested for a mass collection of metabolites that accumulated in these tissues. The GWA results from this data will be useful in discovering candidate genes for each metabolite.

The 440 accessions were grown in a growth room at Purdue University. Leaf and stem tissue were collected and metabolites extracted by the Chapple lab. The total leaf metabolite data comprises of 4,668 metabolites and the stem metabolite data comprises of 3,905 metabolites. The samples were prepared and analyzed by Dr. Li at North Carolina State University.

As an example of the potential of the metabolite GWA results, Li et al. (2014) published a paper using the data of 96 accessions of *A. thaliana* demonstrating the use of

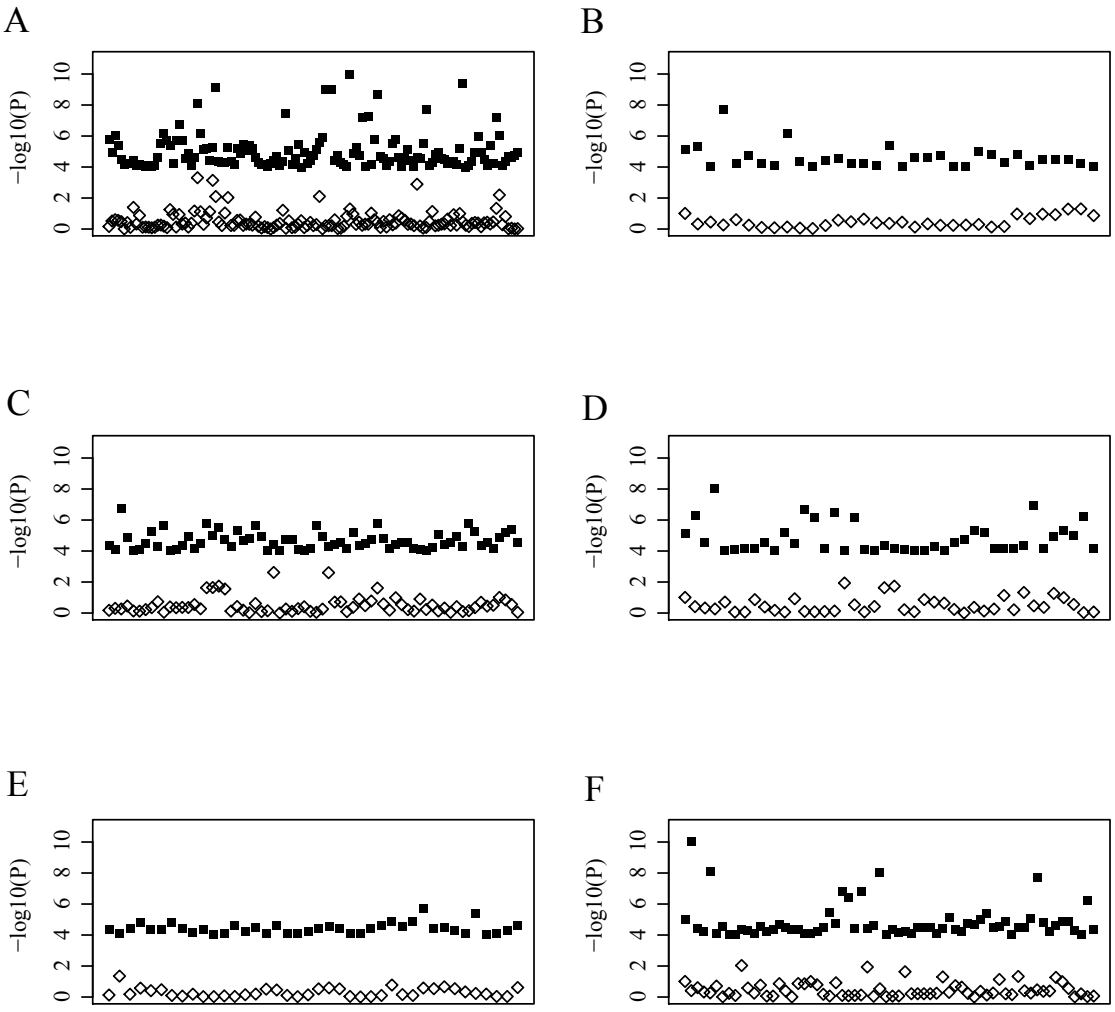

Figure 3.16. Comparisons of SNP *p*-values between seed lethality phenotypes and seed area. The significant SNPs of selected hybrid incompatibility phenotypes were compared to the *p*-values of SNPs calculated for Average Area for seed size using the 211K SNPs EMMAX results. The hybrid incompatibility phenotypes were: A) %P, B) %G, C) %V, D) %PGV, E) %P/%PV F) %S. The y-axis is the *p*-value transformed to the negative log. SNPs are ordered according to position in genome, but the x-axis is not an indication of where in the genome the SNPs are found. The darkened squares are the *p*-values of the significant SNPs for hybrid incompatibility, and the open diamonds are *p*-values of SNPs calculated for Average Area.

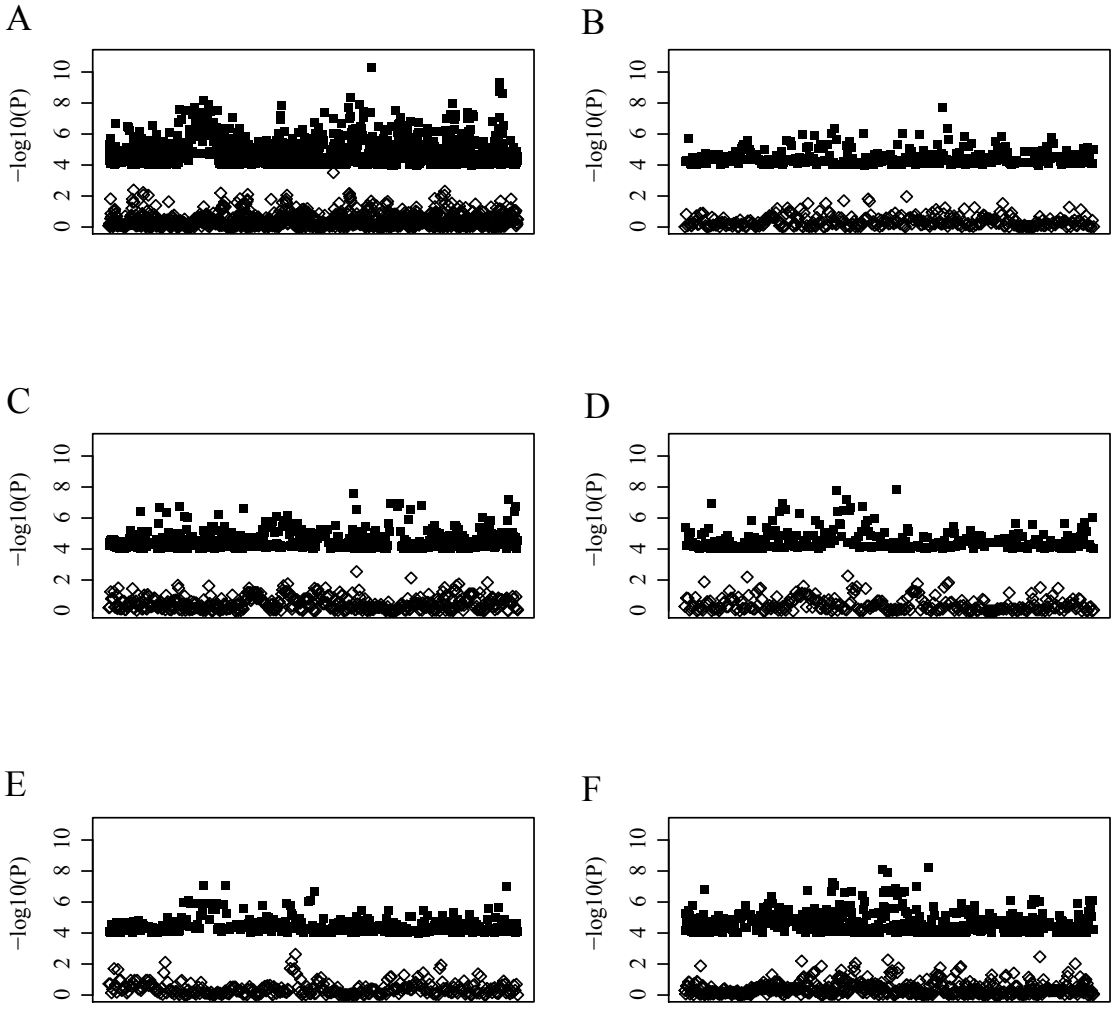

Figure 3.17. Comparisons of SNP *p*-values between seed lethality phenotypes and seed area. The significant SNPs of selected hybrid incompatibility phenotypes were compared to the *p*-values of SNPs calculated for Average Area for seed size using the 1.6M SNPs EMMAX results. The hybrid incompatibility phenotypes were: A) %P, B) %G, C) %V, D) %PGV, E) %P/%PV F) %S. The y-axis is the *p*-value transformed to the negative log. SNPs are ordered according to position in genome, but the x-axis is not an indication of where in the genome the SNPs are found. The darkened squares are the *p*-values of the significant SNPs for hybrid incompatibility, and the open diamonds are *p*-values of SNPs calculated for Average Area.

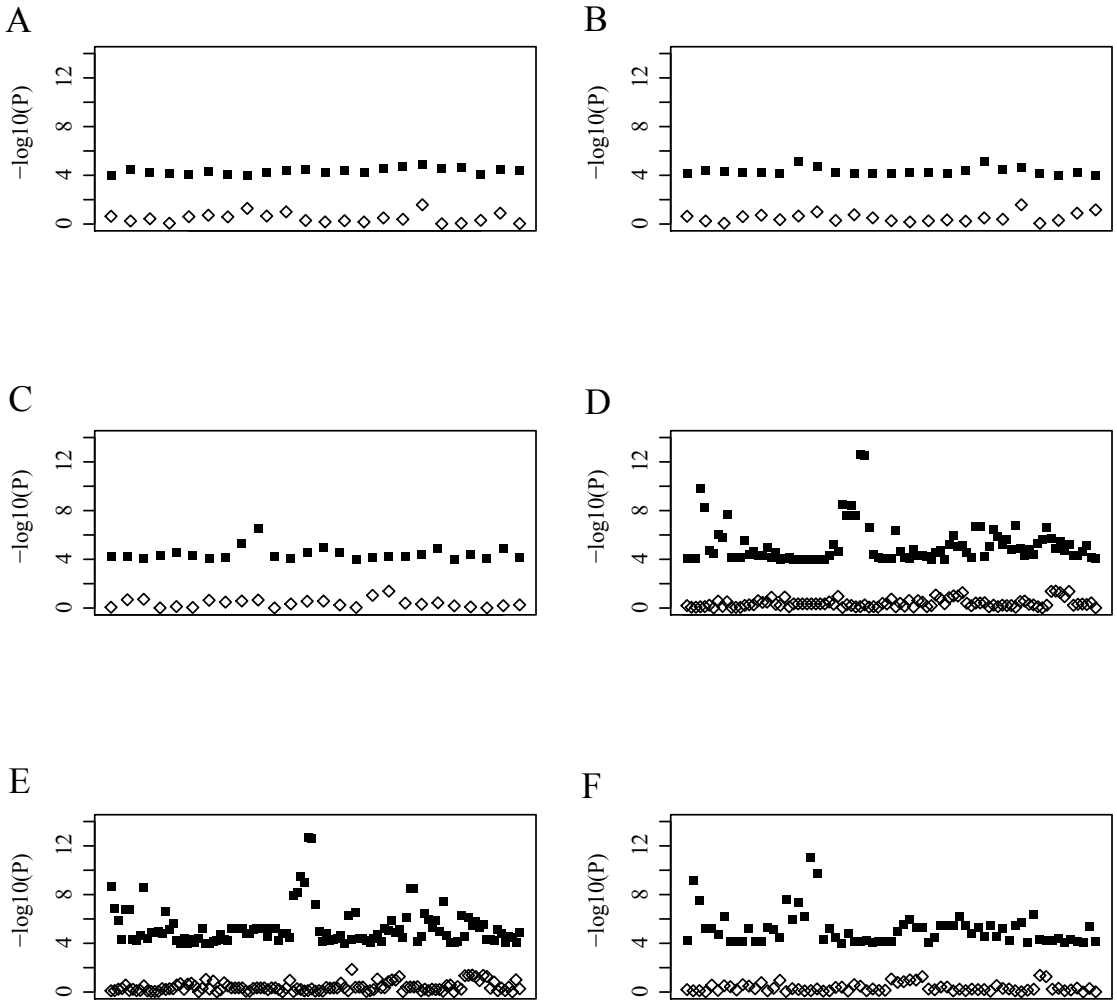

Figure 3.18. Comparisons of SNP *p*-values between seed size phenotypes and %P in hybrid incompatibility. The significant SNPs of selected seed size phenotypes were compared to the *p*-values of SNPs calculated for %P for hybrid incompatibility using the 211K SNPs EMMAX results. The seed size phenotypes were: A) Area, B) Perimeter, C) Circumference, D) MIB, E) MIG, F) MIR.. The y-axis is the *p*-value transformed to the negative log. SNPs are ordered according to position in genome, but the x-axis is not an indication of where in the genome the SNPs are found. The darkened squares are the *p*-values of the significant SNPs for seed size, and the open diamonds are *p*-values of SNPs calculated for %P.

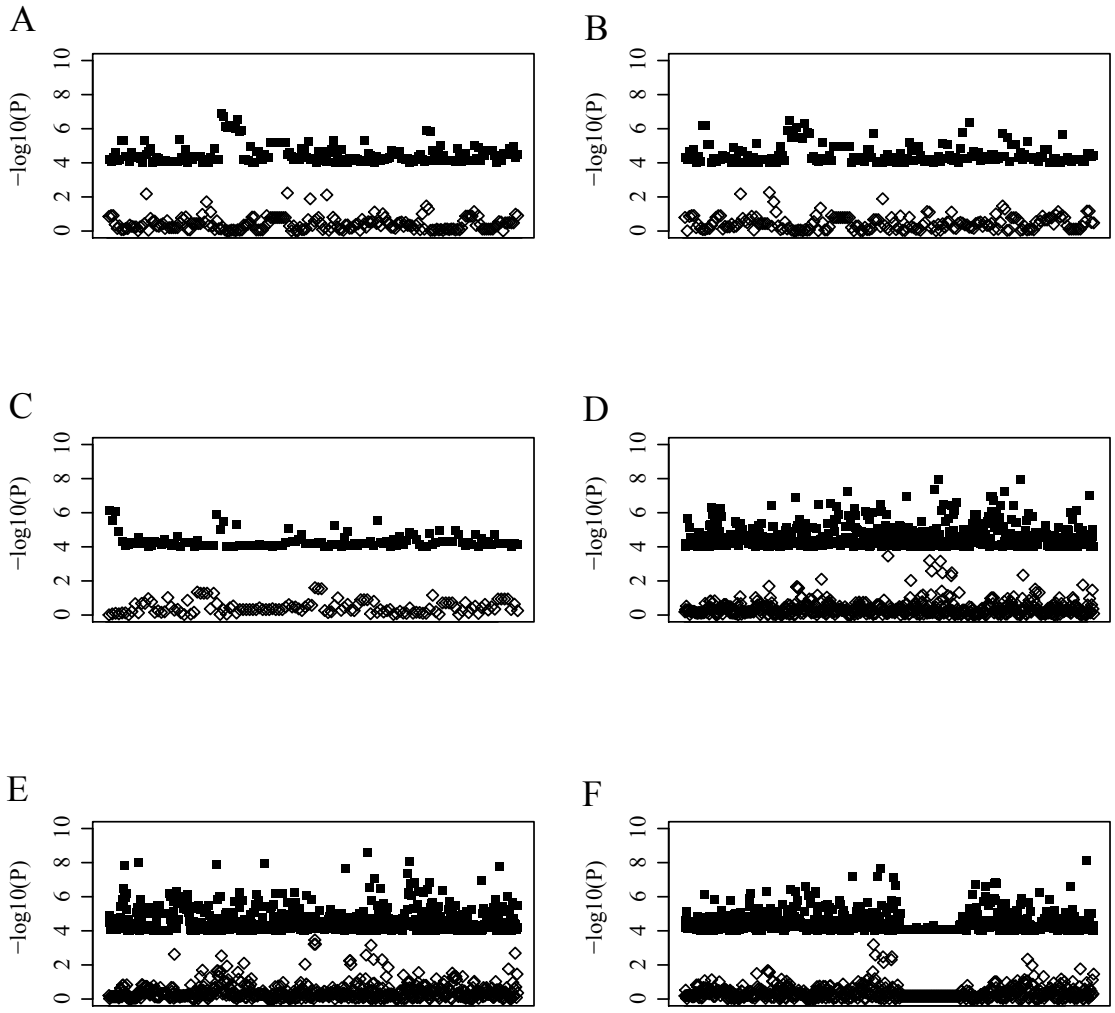

Figure 3.19. Comparisons of SNP *p*-values between seed size phenotypes and %P in hybrid incompatibility. The significant SNPs of selected seed size phenotypes were compared to the *p*-values of SNPs calculated for %P for hybrid incompatibility using the 1.6M SNPs EMMAX results. The seed size phenotypes were: A) Area, B) Perimeter, C) Circumference, D) MIB, E) MIG, F) MIR.. The y-axis is the *p*-value transformed to the negative log. SNPs are ordered according to position in genome, but the x-axis is not an indication of where in the genome the SNPs are found. The darkened squares are the *p*-values of the significant SNPs for seed size, and the open diamonds are *p*-values of SNPs calculated for %P.

GWA to map the gene that is responsible for regulating the synthesis of dihydroxybenzoic acids (Li et al. 2014). Li et al. (2014) used QTL to find loci that were responsible for the synthesis of different dihydroxybenzoic acids. Using my pipeline, the metabolite phenotypes were run using the EMMAX software. The GWA results showed significant SNPS within the QTL locus and the gene linked to the significant SNPs was AT5G03490, a putative *UGT* gene (Li et al. 2014).

As seen with the Li et al. (2014) study, GWA is a great tool for finding new genes involved in metabolite biosynthesis. The information available for the accumulative 8,573 metabolites in the leaf and stem tissue will be an enormous help for understanding metabolite synthesis. The GWA results for the leaf and stem metabolites are available for those that are interested in learning more about the natural variation of metabolites in *A. thaliana.*

3.6 Conclusion

EMMAX and MLMM are both beneficial and disadvantageous statistical models for linking genotype to phenotype. MLMM produces a minimum of significant SNPs to analyze. This makes follow-up experiments easy to decide and plausible. However, interpreting the *p*-values of all other SNPs other than the significant SNPs is difficult since their *p*-values are biased because of the significant SNPs. This reduces the opportunities to learn more about other genes and processes that might be contributing to the phenotype, but the effects are too small to be statistically significant.

The EMMAX results give unbiased *p*-values and so other experiments can be conducted, such as looking at the *p*-values of SNPs linked to candidate genes that do not show up in the putative gene lists. The downfall of EMMAX is the number of significant

SNPs that are produced, which increases with the number of initial SNPs used in the analysis. This could be an indication that adding more SNPs increases the number of false positives, and a new significant cutoff needs to be calculated. The number of significant SNPs is too high to do follow-up experiments when the standard significant cutoff ($\alpha \leq 1x10^{-5}$) is used; therefore, discretion must be used when deciding which genes to explore for future work.

Another aspect that differs between the two methods is the time required for running EMMAX and MLMM. MLMM is much slower, and the computation power required to analyze 5M SNPs is too high and the software crashes. Running 1.6M SNPs is very time consuming, but it is doable, taking several hours per phenotype. EMMAX, on the other hand, is much faster. Increasing the number of SNPs does increase the time required for finishing an analysis using EMMAX, but the time is still manageable, and the computational power is not maxed out.

3.7 References

Abdeen, A, B Miki (2009) The pleiotropic effects of the bar gene and glufosinate on the Arabidopsis transcriptome. Plant Biotechnol J 7:266–282

Adams, S, R Vinkenoog, M Spielman, HG Dickinson, RJ Scott (2000) Parent-of-origin effects on seed development in Arabidopsis thaliana require DNA methylation. Development 127:2493–2502

Araújo, WL, K Ishizaki, A Nunes-Nesi, TR Larson, T Tohge, I Krahnert, S Witt, T Obata, N Schauer, I a Graham, CJ Leaver, AR Fernie (2010) Identification of the 2-hydroxyglutarate and isovaleryl-CoA dehydrogenases as alternative electron donors linking lysine catabolism to the electron transport chain of Arabidopsis mitochondria. Plant Cell 22:1549–63

Atwell, S, YS Huang, BJ Vilhjálmsson, G Willems, M Horton, Y Li, D Meng, A Platt, AM Tarone, TT Hu, R Jiang, NW Muliyati, X Zhang, MA Amer, I Baxter, B Brachi, J Chory, C Dean, M Debieu, J de Meaux, JR Ecker, N Faure, JM Kniskern, JDG Jones, T Michael, A Nemri, F Roux, DE Salt, C Tang, M Todesco, MB Traw, D Weigel, P Marjoram, JO Borevitz, J Bergelson, M Nordborg (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627–31

Autran, D, C Baroux, MT Raissig, T Lenormand, M Wittig, S Grob, A Steimer, M Barann, UC Klostermeier, O Leblanc, J-P Vielle-Calzada, P Rosenstiel, D Grimanelli, U Grossniklaus (2011) Maternal epigenetic pathways control parental contributions to Arabidopsis early embryogenesis. Cell 145:707–19

Bagni, N, A Tassoni (2001) Biosynthesis, oxidation and conjugation of aliphatic polyamines in higher plants. Amino Acids 20:301–317

Baroux, C, D Autran, CS Gillmor, D Grimanelli, U Grossniklaus (2008) The Maternal to Zygotic Transition in Animals and Plants. Cold Spring Harb Symp Quant Biol 73:89–100

Baroux, C, R Blanvillain, P Gallois (2001) Paternally inherited transgenes are down-regulated but retain low activity during early embryogenesis in Arabidopsis. FEBS Lett 509:11–16

Bauwe, H, M Hagemann, AR Fernie (2010) Photorespiration: players, partners and origin. Trends Plant Sci 15:330–336

Berger, F, PE Grini, A Schnittger (2006) Endosperm: an integrator of seed growth and development. Curr Opin Plant Biol 9:664–70

Broz, AK, DK Manter, RM Callaway, MW Paschke, JM Vivanco (2008) A molecular approach to understanding plant–plant interactions in the context of invasion biology. Funct Plant Biol 35:1123

Bukovac, MJ, PD Petracek (1993) Characterizing Pesticide and Surfactant Penetration with Isolated Plant Cuticles. Pestic Sci 37:179–194

Burkart-Waco, D, C Josefsson, B Dilkes, N Kozloff, O Torjek, R Meyer, T Altmann, L Comai (2012) Hybrid incompatibility in Arabidopsis is determined by a multiple-locus genetic network. Plant Physiol 158:801–12

Burkart-Waco, D, K Ngo, B Dilkes, C Josefsson, L Comai (2013) Early disruption of maternal-zygotic interaction and activation of defense-like responses in Arabidopsis interspecific crosses. Plant Cell 25:2037–55

Bushell, C, M Spielman, RJ Scott (2003) The Basis of Natural and Artificial Postzygotic Hybridization Barriers in Arabidopsis Species. Plant Cell 15:1430–1442

Cao, J, K Schneeberger, S Ossowski, T Günther, S Bender, J Fitz, D Koenig, C Lanz, O Stegle, C Lippert, X Wang, F Ott, J Müller, C Alonso-Blanco, K Borgwardt, KJ Schmid, D Weigel (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43:956–63

Chan, EKF, HC Rowe, DJ Kliebenstein (2010) Understanding the Evolution of Defense Metabolites in Arabidopsis thaliana Using Genome-wide Association Mapping. Genetics 185:991–1007

Chang, PL, BP Dilkes, M McMahon, L Comai, S V Nuzhdin (2010) Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol 11:R125

Chaudhury, AM, L Ming, C Miller, S Craig, ES Dennis, WJ Peacock (1997) Fertilization-independent seed development in Arabidopsis thaliana. Proc Natl Acad Sci U S A 94:4223–4228

Chen, F, D Tholl, JC D'Auria, A Farooq, E Pichersky, J Gershenzon (2003) Biosynthesis and Emission of Terpenoid Volatiles from Arabidopsis Flowers. Plant Cell 15:481–494

Chen, M, M Ha, E Lackey, J Wang, ZJ Chen (2008) RNAi of met1 reduces DNA methylation and induces genome-specific changes in gene expression and centromeric small RNA accumulation in Arabidopsis allopolyploids. Genetics 178:1845–58

Chen, Y, Q-Y Pang, Y He, N Zhu, I Branstrom, X-F Yan, S Chen (2012) Proteomics and metabolomics of Arabidopsis responses to perturbation of glucosinolate biosynthesis. Mol Plant 5:1138–50

Comai, L, a P Tyagi, K Winter, R Holmes-Davis, SH Reynolds, Y Stevens, B Byers (2000) Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. Plant Cell 12:1551–68

Coschigano, KT, R Melo-Oliveira, J Lim, GM Coruzzi (1998) Arabidopsis gls Mutants and Distinct Fd-GOGAT Genes : Implications for Photorespiration and Primary Nitrogen Assimilation. Plant Cell 10:741–752

D'Auria, JC, J Gershenzon (2005) The secondary metabolism of Arabidopsis thaliana: growing like a weed. Curr Opin Plant Biol 8:308–16

Däschner, K, I Couée, S Binder (2001) The mitochondrial isovaleryl-coenzyme a dehydrogenase of arabidopsis oxidizes intermediates of leucine and valine catabolism. Plant Physiol 126:601–12

Däschner, K, C Thalheim, C Guha, a Brennicke, S Binder (1999) In plants a putative isovaleryl-CoA-dehydrogenase is located in mitochondria. Plant Mol Biol 39:1275–82

Dilkes, BP, M Spielman, R Weizbauer, B Watson, D Burkart-Waco, RJ Scott, L Comai (2008) The maternally expressed WRKY transcription factor TTG2 controls lethality in interploidy crosses of Arabidopsis. PLoS Biol 6:2707–20

Dixon, R a., D Strack (2003) Phytochemistry meets genome analysis, and beyond..... Phytochemistry 62:815–816

Erilova, A, L Brownfield, V Exner, M Rosa, D Twell, O Mittelsten Scheid, L Hennig, C Köhler (2009) Imprinting of the polycomb group gene MEDEA serves as a ploidy sensor in Arabidopsis. PLoS Genet 5:e1000663

Filiault, DL, JN Maloof (2012) A genome-wide association study identifies variants underlying the Arabidopsis thaliana shade avoidance response. PLoS Genet 8:e1002589

Fournier-Level, a, a Korte, MD Cooper, M Nordborg, J Schmitt, a M Wilczek (2011) A map of local adaptation in Arabidopsis thaliana. Science 334:86–9

Foy, C (1964) Foliar Penetration: Review of Herbicide Penetration through Plant Surfaces. J Agric Food Chem 12:473–476

Garcia, D, JNF Gerald, F Berger (2005) Maternal Control of Integument Cell Elongation and Zygotic Control of Endosperm Growth Are Coordinated to Determine Seed Size in Arabidopsis. Plant Cell 17:52–60

Garcia, D, V Saingery, P Chambrier, U Mayer, G Jürgens, F Berger (2003) Arabidopsis haiku Mutants Reveal New Controls of Seed Size by Endosperm. Plant Physiol 131:1661–1670

Gardner, SN, J Gressel, M Mangel (1998) A revolving dose strategy to delay the evolution of both quantitative vs major monogene resistance to pesticides and drugs 44:161–180

Givan, C V, KW Joy, L a Kleczkowski (1988) A decade of photorespiratory nitrogen cycling. Trends Biochem Sci 13:433–7

Good, AG, SJ Johnson, M De Pauw, RT Carroll, N Savidov, J Vidmar, Z Lu, G Taylor, V Stroeher (2007) Engineering nitrogen use efficiency with alanine aminotransferase. Can J Bot 85:252–262

Grimanelli, D, E Perotti, J Ramirez, O Leblanc (2005) Timing of the Maternal-to-Zygotic Transition during Early Seed Development in Maize. Plant Cell 17:1061–1072

Grubb, CD, S Abel (2006) Glucosinolate metabolism and its control. Trends Plant Sci 11:89–100

Gu, L, a D Jones, RL Last (2010) Broad connections in the Arabidopsis seed metabolic network revealed by metabolite profiling of an amino acid catabolism mutant. Plant J 61:579–90

Guitton, A-E, F Berger (2005) Control of reproduction by Polycomb Group complexes in animals and plants. Int J Dev Biol 49:707–16

Gupta, G, B Grund, R Narayanan (1991) Photosynthesis and nitrogenase activity in soybean treated with sulphur dioxide and molybdenum. Plant Sci 79:157–161

Haig, D, M Westoby (1989) Parent-Specific Gene Expression and the Triploid Endosperm. Am Nat 134:147–155

Haig, D, M Westoby (1991) Genomic imprinting in endosperm : its effect on seed development in crosses between species , and between different ploidies of the same species , and its implications for the evolution of apomixis. Philos Trans Biol Sci 333:1–13

Hartmann, T (2007) From waste products to ecochemicals: fifty years research of plant secondary metabolism. Phytochemistry 68:2831–46

Häusler, RE, RD Blackwelf, PJ Lea, RC Leegood (1994) Control of photosynthesis in barley leaves with reduced activities of glutamine synthetase or glutamate synthase I . Plant characteristics and changes in nitrate , ammonium and amino acids. Planta 194:406–417

Heap, I (2015) The International Survey of Herbicide Resistant Weeds. www.weedscience.com

Hehenberger, E, D Kradolfer, C Köhler (2012) Endosperm cellularization defines an important developmental transition for embryo development. Development 139:2031–9

Henry, IM, BP Dilkes, L Comai (2007) Genetic basis for dosage sensitivity in Arabidopsis thaliana. PLoS Genet 3:e70

Henry, IM, BP Dilkes, K Young, B Watson, H Wu, L Comai (2005) Aneuploidy and genetic variation in the Arabidopsis thaliana triploid response. Genetics 170:1979–88

Horton, MW, N Bodenhausen, K Beilsmith, D Meng, BD Muegge, S Subramanian, MM Vetter, BJ Vilhjálmsson, M Nordborg, JI Gordon, J Bergelson (2014) Genome-wide association study of Arabidopsis thaliana leaf microbial community. Nat Commun 5:5320

Horton, MW, AM Hancock, YS Huang, C Toomajian, S Atwell, A Auton, NW Muliyati, A Platt, FG Sperone, BJ Vilhjálmsson, M Nordborg, JO Borevitz, J Bergelson (2012) Genome-wid patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel 44:212–216

Ide, Y, M Kusano, A Oikawa, A Fukushima, H Tomatsu, K Saito, MY Hirai, T Fujiwara (2011) Effects of molybdenum deficiency and defects in molybdate transporter MOT1 on transcript accumulation and nitrogen/sulphur metabolism in Arabidopsis thaliana. J Exp Bot 62:1483–97

Josefsson, C, B Dilkes, L Comai (2006) Parent-dependent loss of gene silencing during interspecies hybridization. Curr Biol 16:1322–8

Kamm, A, I Galasso, T Schmidt, JS Heslop-Harrison (1995) Analysis of a repetitive DNA family from Arabidopsis arenosa and relationships between Arabidopsis species. Plant Mol Biol 27:853–862

Kasukabe, Y, L He, K Nada, S Misawa, I Ihara, S Tachibana (2004) Overexpression of spermidine synthase enhances tolerance to multiple environmental stresses and up-regulates the expression of various stress-regulated genes in transgenic Arabidopsis thaliana. Plant Cell Physiol 45:712–22

Kaundun, SS (2010) An aspartate to glycine change in the carboxyl transferase domain of acetyl CoA carboxylase and non-target-site mechanism(s) confer resistance to ACCase inhibitor herbicides in a Lolium multiflorum population. Pest Manag Sci 66:1249–56

Keys, AJ, IF Bird, MJ Cornelius, PJ Lea, RM Wallsgove, BJ Miflin (1978) Photorespiratory nitrogen cycle. Nature 275:741–743

Kliebenstein, DJ, J Kroymann, P Brown, A Figuth, D Pedersen, J Gershenzon, T Mitchell-olds (2001a) Genetic Control of Natural Variation in Arabidopsis Glucosinolate Accumulation. Plant Physiol 126:811–825

Kliebenstein, DJ, VM Lambrix, M Reichelt, J Gershenzon, T Mitchell-olds (2001b) Gene Duplication in the Diversification of Secondary Metabolism : Tandem 2-Oxoglutarate–Dependent Dioxygenases Control Glucosinolate Biosynthesis in Arabidopsis. Plant Cell 13:681–693

Koch, MA, B Haubold, T Mitchell-olds (2000) Comparative Evolutionary Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in Arabidopsis, Arabis, and Related Genera (Brassicaceae). Mol Biol Evol 17:1483–1498

Köhler, C, L Hennig, C Spillane, S Pien, W Gruissem, U Grossniklaus (2003) The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene PHERES1. Genes Dev 17:1540–53

Koornneef, M, C Alonso-Blanco, D Vreugdenhil (2004) Naturally occurring genetic variation in Arabidopsis thaliana. Annu Rev Plant Biol 55:141–72

Kradolfer, D, L Hennig, C Köhler (2013a) Increased maternal genome dosage bypasses the requirement of the FIS polycomb repressive complex 2 in Arabidopsis seed development. PLoS Genet 9:e1003163

Kradolfer, D, P Wolff, H Jiang, A Siretskiy, C Köhler (2013b) An imprinted gene underlies postzygotic reproductive isolation in Arabidopsis thaliana. Dev Cell 26:525–35

Kroymann, J, S Textor, JG Tokuhisa, KL Falk, J Gershenzon, T Mitchell-olds (2001) A Gene Controlling Variation in Arabidopsis Glucosinolate Composition Is Part of the Methionine Chain Elongation Pathway. Plant Physiol 127:1077–1088

Kuittinen, H, M Aguadé (2000) Nucleotide Variation at the CHALCONE ISOMERASE Locus in Arabidopsis thaliana. Genetics 155:863–872

Lacuesta, M, B Gonzalezmoro, C Gonzalezmurua, A Muozrueda (1990) Temporal study of the effect of phosphinothricin on the activity of glutamine-synthetase, glutamate-dehydrogenase and nitrate reductase in Medicago sativa L. J Plant Physiol 1364:410–414

Leegood, RC, PJ Lea, MD Adcock, RE Häusler (1995) The regulation and control of photorespiration. J Exp Bot 46:1397–1414

Li, X, E Svedin, H Mo, S Atwell, BP Dilkes, C Chapple (2014) Exploiting Natural Variation of Secondary Metabolism Identifies a Gene Controlling the Glycosylation Diversity of Dihydroxybenzoic Acids in Arabidopsis thaliana. Genetics 198:1267–76

Li, Y, F Beisson, AJK Koo, I Molina, M Pollard, J Ohlrogge (2007) Identification of acyltransferases required for cutin biosynthesis and production of cutin with suberin-like monomers. Proc Natl Acad Sci U S A 104:18339–44

Li, Y, Y Huang, J Bergelson, M Nordborg, JO Borevitz (2010) Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. Proc Natl Acad Sci U S A 107:21199–204

Liepman, AH, LJ Olsen (2003) Alanine Aminotransferase Homologs Catalyze the Glutamate : Glyoxylate Aminotransferase Reaction in Peroxisomes of Arabidopsis. Plant Physiol 131:215–227

Liu, H, C Hu, X Sun, Q Tan, Z Nie, X Hu (2009) Interactive effects of molybdenum and phosphorus fertilizers on photosynthetic characteristics of seedlings and grain yield of Brassica napus. Plant Soil 326:345–353

Luo, M, ES Dennis, F Berger, WJ Peacock, A Chaudhury (2005) MINISEED3 (MINI3), a WRKY family gene, and HAIKU (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in Arabidopsis. Proc Natl Acad Sci U S A 102:17531–17536

Manderscheid, R, A Wild (1986) Studies on the mechanism of inhibition of phosphinothricin of Glutamine synthetase isolated from Triticum aestivum L. J Plant Physiol 123:135–142

McClung, CR, M Hsu, JE Painter, JM Gagne, SD Karlsberg, P a Salomé (2000) Integrated temporal regulation of the photorespiratory pathway. Circadian regulation of two Arabidopsis genes encoding serine hydroxymethyltransferase. Plant Physiol 123:381–92

Mendel, RR (2002) Molybdoenzymes and molybdenum cofactor in plants. J Exp Bot 53:1689–1698

Miflin, BJ, PJ Lea (1980) Ammonia assimilation. Pages 169–202 *in* BJ Miflin, ed. The Biochemistry of PlantsVol 5. New York: Academic Press

Morris, PF, DB Layzell, DT Canvin (1989) Photorespiratory ammonia does not inhibit photosynthesis in glutamate synthase mutants of Arabidopsis. Plant Physiol 89:498–500

Nodine, MD, DP Bartel (2012) Maternal and paternal genomes contribute equally to the transcriptome of early plant embryos. Nature 482:94–7

Novitskaya, L, SJ Trevanion, S Driscoll, CH Foyer, G Noctor (2002) How does photorespiration modulate leaf amino acid contents? A dual approach through modelling and metabolite analysis. Plant, Cell Environ 25:821–835

Nowack, MK, A Ungru, KN Bjerkan, PE Grini, A Schnittger (2010) Reproductive cross-talk: seed development in flowering plants. Biochem Soc Trans 38:604–12

O'Kane, SL, BA Schaal, IA Al-shehbaz (1996) The Origins of Arabidopsis suecica (Brassicaceae) as Indicated by Nuclear rDNA Sequences. Syst Bot 21:559–566

Peterhansel, C, I Horst, M Niessen, C Blume, R Kebeish, S Kürkcüoglu, F Kreuzaler (2010) Photorespiration. Page e0130 *in* Arabidopsis Book. 8th ed. Rockville: The American Society of Plant Biologists

Platt, A, M Horton, YS Huang, Y Li, AE Anastasio, NW Mulyati, J Agren, O Bossdorf, D Byers, K Donohue, M Dunning, EB Holub, A Hudson, V Le Corre, O Loudet, F Roux, N Warthmann, D Weigel, L Rivero, R Scholl, M Nordborg, J Bergelson, JO Borevitz (2010) The scale of population structure in Arabidopsis thaliana. PLoS Genet 6:e1000843

Potel, F, M-H Valadier, S Ferrario-Méry, O Grandjean, H Morin, L Gaufichon, S Boutet-Mercey, J Lothier, SJ Rothstein, N Hirose, A Suzuki (2009) Assimilation of excess ammonium into amino acids and nitrogen translocation in Arabidopsis thaliana--roles of glutamate synthases and carbamoylphosphate synthetase in leaves. FEBS J 276:4061–76

R Core Team (2013) R: A language and environment for statistical computing. Vienna, Austria

Robertson, J, K Farnden (1980) Ultrastructure and metabolism of the developing legume root nodule. Pages 65–113 *in* BJ Miflin, ed. The Biochemistry of PlantsVol 5. New York: Academic Press

Ruegger, M, C Chapple (2001) Mutations That Reduce Sinapoylmalate Accumulation in Arabidopsis thaliana Define Loci With Diverse Roles in Phenylpropanoid Metabolism. Genetics 159:1741–1749

Schatlowski, N, P Wolff, J Santos-González, V Schoft, A Siretskiy, R Scott, H Tamaru, C Köhler (2014) Hypomethylated pollen bypasses the interploidy hybridization barrier in Arabidopsis. Plant Cell 26:3556–68

Scott, RJ, M Spielman, J Bailey, HG Dickinson (1998) Parent-of-origin effects on seed development in Arabidopsis thaliana. Develop 125:3329–3341

Seabra, AR, P a Pereira, JD Becker, HG Carvalho (2012) Inhibition of glutamine synthetase by phosphinothricin leads to transcriptome reprograming in root nodules of Medicago truncatula. Mol Plant Microbe Interact 25:976–92

Shirley, BW, WL Kubasek, G Stor, E Bruggemann, M Koornneef, FM Ausubel, HM Goodman (1995) Analysis of Arabidopsis mutants deficient in flavonoid biosynthesis. Plant J 8:659–671

Smith, CC, SD Fretwell (1974) The Optimal Balance between Size and Number of Offspring. Am Nat 108:499–506

Somerville, SC, WL Ogren (1981) Photorespiration-Deficient Mutants of Arabidopsis thaliana Lacking Mitochondrial Serine Transhydroxymethylase Activity. Plant Physiol 67:666–671

Somerville, SC, WL Ogren (1983) An Arabidopsis thaliana mutant defective in chloroplast dicarboxylate transport. Proc Natl Acad Sci U S A 80:1290–4

Sønderby, IE, F Geu-flores, BA Halkier (2010) Biosynthesis of glucosinolates – gene discovery and beyond. Trends Plant Sci 15:283–290

Stallmeyer, B, G Schwarz, J Schulze, a Nerlich, J Reiss, J Kirsch, RR Mendel (1999) The neurotransmitter receptor-anchoring protein gephyrin reconstitutes molybdenum cofactor biosynthesis in bacteria, plants, and mammalian cells. Proc Natl Acad Sci U S A 96:1333–8

Suzuki, A, DB Knaff (2005) Glutamate synthase: structural, mechanistic and regulatory properties, and role in the amino acid metabolism. Photosynth Res 83:191–217

Ta, TC, KW Joy (1986) Metabolism of some amino acids in relation to the photorespiratory nitrogen cycle of pea leaves. Planta 169:117–22

Tholl, D, F Chen, J Petri, J Gershenzon, E Pichersky (2005) Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from Arabidopsis flowers. Plant J 42:757–71

Tsukagoshi, H, T Saijo, D Shibata, A Morikami, K Nakamura (2005) Analysis of a Sugar Response Mutant of Arabidopsis Identified a Novel B3 Domain Protein That Functions as an Active Transcriptional Repressor. Plant Physiol 138:675–685

Vielle-Calzada, J-P, R Baskar, U Grossniklaus (2000) Delayed activation of the paternal genome during seed development. Nature 404:91–94

Voll, LM, A Jamai, P Renné, H Voll, CR Mcclung, APM Weber, P Renne (2006) The Photorespiratory Arabidopsis shm1 Mutant is Deficient in SHM1 140:59–66

Walia, H, C Josefsson, B Dilkes, R Kirkbride, J Harada, L Comai (2009) Dosage-dependent deregulation of an AGAMOUS-LIKE gene cluster contributes to interspecific incompatibility. Curr Biol 19:1128–32

Walker, KA, C V Givan, AJ Keys (1984) Glutamic acid metabolism and the photorespiratory nitrogen cycle in wheat leaves: Metabolic consequences of elevated ammonia concentratiosn and of blockign ammonia assimilation. Plant Physiol 75:60–66

Wallsgrove, RM, AC Kendall, NP Hall, JC Turner, PJ Lea (1986) Carbon and nitrogen metabolism in a barley (Hordeum vulgare L.) mutant with impaired chloroplast dicarboxylate transport. Planta 168:324–329

Wallsgrove, RM, AJ Keys, PJ Eea, BJ Miflin (1983) Photosynthesis , photorespiration and nitrogen metabolism. Plant, Cell Environ 6:301–309

Weigel, D (2012) Natural variation in Arabidopsis: from molecular genetics to ecological genomics. Plant Physiol 158:2–22

Wendler, C, M Barniske, a Wild (1990) Effect of phosphinothricin (glufosinate) on photosynthesis and photorespiration of C3 and C 4 plants. Photosynth Res 24:55–61

Woo, KC, JF Morot-Gaudry, RE Summons, OC B (1982) Evidence for the Glutamine Synthetase/Glutamate Synthase Pathway during the Photorespiratory Nitrogen Cycle in Spinach Leaves. Plant Physiol 70:1514–7

Xiang, D, P Venglat, C Tibiche, H Yang, E Risseeuw, Y Cao, V Babic, M Cloutier, W Keller, E Wang, G Selvaraj, R Datla (2011) Genome-wide analysis reveals gene expression and metabolic network dynamics during embryo development in Arabidopsis. Plant Physiol 156:346–56

Yamaguchi, K, Y Takahashi, T Berberich, A Imai, T Takahashi, AJ Michael, T Kusano (2007) A protective role for the polyamine spermine against drought stress in Arabidopsis. Biochem Biophys Res Commun 352:486–90

Yonekura-Sakakibara, K, K Saito (2009) Functional genomics for plant natural product biosynthesis. Nat Prod Rep 26:1466–87

Yu, Q, I Abdallah, H Han, M Owen, S Powles (2009) Distinct non-target site mechanisms endow resistance to glyphosate, ACCase and ALS-inhibiting herbicides in multiple herbicide-resistant Lolium rigidum. Planta 230:713–23

Yuan, JS, PJ Tranel, CN Stewart (2007) Non-target-site herbicide resistance: a family business. Trends Plant Sci 12:6–13

Zhang, Y, X Hu, Y Shi, Z Zou, F Yan, Y Zhao, H Zhang (2013) Beneficial Role of Exogenous Spermidine on Nitrogen Metabolism in Tomato Seedlings Exposed to Saline-alkaline Stress. J Am Soc Hortic Sci 138:38–49

Zhang, Y, K Sun, FJ Sandoval, K Santiago, S Roje (2010) One-carbon metabolism in plants: characterization of a plastid serine hydroxymethyltransferase. Biochem J 430:97–105

Zhao, C-R, Y Sawaki, N Sakurai, D Shibata, H Koyama (2010) Transcriptomic profiling of major carbon and amino acid metabolism in the roots of Arabidopsis thaliana treated with various rhizotoxic ions. Soil Sci Plant Nutr 56:150–162

APPENDICES

Appendix A R code for hierarchical clustering

```
Bla_5 <- read.csv("File",header=TRUE,na.string="NA")
Bla_5 <-data.matrix(Bla_5)
Bla_5_hclust <- hclust(dist(Bla_5,method="manhattan"),method="complete")
plot(Bla_5_hclust)
pdf("Bla_5.hclust.pdf",width=8,height=11,pointsize=1)
plot(Bla_5_hclust,main="Subtree of Bla-5")
dev.off()

Wa_1 <- read.csv("File",header=TRUE,na.string="NA")
Wa_1 <-data.matrix(Wa_1)
Wa_1_hclust <- hclust(dist(Wa_1,method="manhattan"),method="complete")
plot(Wa_1_hclust)
pdf("Wa_1.hclust.pdf",width=8,height=11,pointsize=1)
plot(Wa_1_hclust,main="Subtree of Wa-1")
dev.off()

M3385S <- read.csv("File",header=TRUE,na.string="NA")
M3385S <-data.matrix(M3385S)
M3385S_hclust <- hclust(dist(M3385S,method="manhattan"),method="complete")
plot(M3385S_hclust)
pdf("M3385S.hclust.pdf",width=8,height=11,pointsize=1)
plot(M3385S_hclust,main="Subtree of M3385S")
dev.off()
```

Appendix B Running_GWAS.pl

```
#! /usr/bin/perl

#-------------------------------------------------------------#
#  This script uses a .csv file of all the phenotypes that need to be analyzed. The first
#  column is the accession name and after that each column is a phenotype. The script
#  takes each column and creates a separate phenoytpe file to be used for EMMAX. The
#  script then runs each phenotype through EMMAX and the output is printed in the
#  directory. For this script to work you need a phenotype file, the software emmax and
#  emmax-kin, a tped file of the SNPS, a tfam file, and  #  the kinship file which can be
#  made using the following code from Kang et al.
#  To run the kinship command: emmax-kin -v -h -s -d 10 [tped_prefix]
#-------------------------------------------------------------#

#! /usr/bin/perl
use strict;

my @trait;                                  #Declare an array
my $total_traits;                           #Declare an element
my $total1;                                 #Declare an element

#Read in the phenotype file
while ( my $line = <> ) {
        chomp $line;

        my @array = split ",", $line;  #Splits each line into an array
        push @trait, \@array;          #An array of arrays.
        $total_traits = $#array;       #Total number of array
}

my $sourcefile;                             #Declare an element

#The following is a loop that will create the phenotype files by going through each row of
each column and printing the phenotype in a new file.
for ( my $t = 1; $t <= $total_traits; $t++ ) {
        my @phenotype;
        push @phenotype, $trait[0][$t];
        push @phenotype, $trait[1][$t];
        $sourcefile = $trait[0][$t];           #$sourcefile=phenotype

        for ( my $g = 1; $g <= $#trait; $g++ ) {
                push @phenotype, $trait[$g][$t];
        }
```

```
        system ("mkdir $sourcefile" );          #Folder named phenotype
        system ( "cd $sourcefile" );            #To open the folder
        system ( "mkfile -nv 100k $sourcefile" );     #Make a file within folder

        #This opens the file in the specific directory and name it [phenotype].phenos
        open ( PHENOTYPE, ">$sourcefile/$sourcefile.phenos" ) || die"$!";

        my $total = 1;
        #The following loop prints the phenotype data into the file. The phenotype
                file needed for EMMAX is created.
        for (my $h = 2; $h <= $#phenotype; $h++ ) {
                print PHENOTYPE "$total 1 $phenotype[$h]\n";
                $total++;
        }

        close(PHENOTYPE);

        #This next script runns EMMAX
        #***User needs to change the name of the tped and tfam file and the kinship
                file. Meaning "428Accessions" and "428Accessions.hIBS.kinf" needs to
        be changed to their correct names.***
        system ( "./emmax -v -d 10 -t 428Accessions -p
                $sourcefile/$sourcefile.phenos -k 428Accessions.hIBS.kinf -o
                $sourcefile/$sourcefile" );

#When the file is closed and the loop starts over it goes back to the original directory and
starts all over, that is so awesome!
}
#--------------------------------The End----------------------#

print "This should be the end!\n";
#This script lets you know that the script has successfullycompleted all the phenotypes in
the file.
```

Appendix C ManhanFiles_Plots.pl

```
#! /usr/bin/perl

#-----------------------------------------------------------------------------------------------#
This script reads all .ps output files from EMMAX and converts them to .manhan files
that are readable to make manhattan plots in R. It also prints an R script for loading
the .manhan.csv file into R and to make a jpeg or pdf of the manhattan plot. The R script
is printed into its own file so that you can just open it and load it into R.
*This script needs to be in the same directory as the EMMAX output folders.
#-----------------------------------------------------------------------------------------------#

use strict;

#-----------------------------------------------------------------------------------------------#
PART A: Creating .manhan files. All folders and files in the directory are saved in an
array. The files and folders that are not from the EMMAX output need to excluded from
the further manipulation, and this is done using conditional statements.  The .ps files
are opened and columns are named and rearranged to follow the file format required for
making manhattan plots in R. The new files are named phenotype.manhan.csv. These
files are put in a new folder called "manhattanfiles".
**The only things that need to be changed in the script is the exclusion of files/folders.
#-----------------------------------------------------------------------------------------------#
my $dir_list = `ls`;                    #`ls` gets all the files from the directory; the files
                                        have a hard return after them!
my @folders = split "\n", $dir_list;    #splits the array using the hard return as
                                        the separator.

system ( "mkdir manhattanfiles" );              #makes a folder called manhattanfiles

#The following loop is going to go through the array @folders, which contains the names
of the folders. It will exclude any of the folders/files specified in the "elsif" conditions
because they do not contain .ps files and kill the script if not excluded. Depending on
your directory you can delete lines or add as many conditions as you need to only include
the EMMAX result folders.

for ( my $h = 0; $h <= $#folders; $h++ ) {
        my @manhan_array;
        if ( $folders[$h] =~ m/ManhanFiles/ ) {                 #Excluding
        } elsif ( $folders[$h] =~ m/manhattanfiles/ ) {         #Excluding
        } elsif ( $folders[$h] =~ m/Aarenosa/ ) {               #Excluding
        } elsif ( $folders[$h] =~ m/SeedSize/ ) {               #Excluding
        } elsif ( $folders[$h] =~ m/Running/ ) {                #Excluding
        } elsif ( $folders[$h] =~ m/428Accessions/ ) {          #Excluding
```

```
	} elsif ( $folders[$h] =~ m/emmax/ ) {		#Excluding
	} elsif ( $folders[$h] =~ m/Significant/ ) {	#Excluding
	} elsif ( $folders[$h] =~ m/Finding/ ) {	#Excluding
	} elsif ( $folders[$h] =~ m/Counting/ ) {	#Excluding
	} elsif ( $folders[$h] =~ m/leaf/ ) {		#Excluding
	} else {				#Will open .ps files of EMMAX
output files
		my $folder = "$folders[$h]";
		my $file = "$folders[$h].ps";
		print "$folder/$file\n";		#Prints which file it is opening.
		#Open the .ps file that I want.
		open ( PVALUES, "$folders[$h]/$folders[$h].ps" ) || die;
	#-------------------------------------------------------------------------------------#
	The following is the script that splices the .ps file into different elements and
	saves them into an array to be used later in this script.
	#-------------------------------------------------------------------------------------#
		while ( my $line = <PVALUES> ) {	#Reads the .ps files
			chomp $line;

			my @array = split "\t", $line; #Splitting the line
			#Splitting the SNPID into different parts to get the
		chromosome and SNP position
			my @chrom = split "", $array[0];
			#Declaring the SNP base pair
			my $bp = substr( $array[0], 1 ); # ",", $array[0], 2;
			my @manhattan;
			$manhattan[0] = $array[0];		#snpid
			$manhattan[1] = $chrom[0];	#chromosome
			$manhattan[2] = $bp;		#basepair
			$manhattan[3] = $array[2];		#p-values
			$manhattan[4] = $array[1];		#beta values
			#Create an array of arrays to be used later on for printing new
			file
			push @manhan_array, \@manhattan;
		}
		close( PVALUES );			#Closing the .ps file
		#-------------------------------------------------------------------------------------#
		This part makes a new file named phentoype.manhan.csv. Printed in
		this file is the file format of the pvalues that is needed to read the file
		into R and to make manhattan plots. The columns are contain the SNP
		ID (SNP), chromosome (CHR), base pair (BP), p-value (P), and beta-
		value (B).
		#-------------------------------------------------------------------------------------#
		#Opens the folder "manhattanfiles" to put in new files
```

```
                system ( "cd manhattanfiles" );
                #makes a file within "manhattanfiles"
                system ( "mkfile -nv 100k manhattanfiles" );
                #opens the file I just made and names it

                open (FILE, ">manhattanfiles/$folders[$h].manhan.csv" ) || die;
                #prints "SNP CHR BP P B" on the file
                        print FILE "SNP CHR BP P B\n";
                        #Loop through array to print EMMAX results
                        for ( my $r = 0; $r <= $#manhan_array; $r++ ) {
                                print FILE "$manhan_array[$r][0]
                                        $manhan_array[$r][1] $manhan_array[$r][2]
                                        $manhan_array[$r][3] $manhan_array[$r][4]\n";
                        }
                close ( FILE );                         # Closes the .ps file
        }
}
#The end of my for loop that goes through each folder containing emmax results. All
the .manhan files are created and are sitting in the folder titled "manhattanfiles"

#-------------------------------------------------------------------------------------------#
PART B: This next script writes an R script to create manhattan plots. The script is
printed in the "manhattanfiles" folder also. The file named "readmanhanfilesintoR.r" will
contain the R script needed to read the .manhan file into R, make a manhattan plot, and
also save the manhattan plot as a jpg/pdf. For example if you have 25 phenotypes you
will have 25 small R scripts that are the same expect that each R script is specific for each
phenotype.
**The only things that need to be changed is the home directory of the final location of
the "manhattanfiles" folder.  If the user does not know this or it changes you can use
Find/Replace later on to change the home directory in the R script. Also, the user needs to
specify if jpeg or pdf should be printed.
NOTE: If you know how to use R for this part, by all means, you can write your own R
script reads the files in and creates the manhattan plots. This was how I knew how to do it.
#-------------------------------------------------------------------------------------------#
#Create file in the folder "manhattanfiles"
system ( "mkfile -nv 100k manhattanfiles" );
#Name file "readmanhanfilesintoR.r"
open (RSCRIPT, ">manhattanfiles/readmanhanfilesintoR.r");

my $total_lines = 1;

#The following script will once again exclude any folder or file that is not EMMAX
output.
for ( my $f = 0; $f <= $#folders; $f++ ) {
        if ( $folders[$f] =~ m/ManhanFiles/ ) {                 #Excluding
```

```
	} elsif ( $folders[$f] =~ m/manhattanfiles/ ) {		#Excluding
	} elsif ( $folders[$f] =~ m/SeedSize/ ) {		#Excluding
	} elsif ( $folders[$f] =~ m/Running/ ) {		#Excluding
	} elsif ( $folders[$f] =~ m/428Accessions/ ) {		#Excluding
	} elsif ( $folders[$f] =~ m/emmax/ ) {			#Excluding
	} elsif ( $folders[$f] =~ m/SignificantSNPs/ ) {	#Excluding
	} elsif ( $folders[$f] =~ m/FindingSignificnatSNPs/ ) {	#Excluding
	} else {						#EMMAX output
		#The R script.
		#**Need to change the path for .manhan.csv files
		print RSCRIPT "$folders[$f] <-  read.table(\"/path/manhattanfiles/
	$folders[$f].manhan.csv\",header=TRUE,na.string=\"NA\")\n\n";
		if ( $total_lines == 1 ) {
			print RSCRIPT "manhattan($folders[$f]);\n\n";
			$total_lines++;
		}
		#To create a pdf instead of jpg change jpg/jpeg to pdf. To create jpg
		instead of pdf change .pdf to .jpg and pdf to jpeg
		print RSCRIPT "dev.print(device=postscript, \"$folders[$f].jpg\",
		onefile=FALSE,horizontal=FALSE);\njpeg(\"$folders[$f].jpg\");\n
		manhattan($folders[$f],pch=20,main=\"$folders[$f]\");\n
	dev.off()\n\n";
	}
}
close (RSCRIPT);
#----------The End----------#
```

Appendix D FindingSignificantSNPs.pl

```
#!/usr/bin/perl

#-----------------------------------------------------------------------------------------------#
This script needs to be saved in a directory that contains the phenotype folders produced
from EMMAX. It will open each folder and find the .ps files. This script reads the .ps file,
which contains the pvalues and then pulls out the SNPs that have a pvalue that is equal to
or lower than what the user specifies. It is then printed into a file that is named
phenotype.sigsnps. This script also adds the allele frequency of each SNP to the file so
you know how common the SNP is. It is the allele frequency of the non-Columbia allele.
#-----------------------------------------------------------------------------------------------#

use strict;

#-----------------------------------------------------------------------------------------------#
PART A: Reading the allele frequencies into the script. THe first line is of the allele
frequency file is ommitted so that the first SNP in the arrays match the first SNP in the
arrays created using the .ps files. This is important for increasing speed of this scrip.
 #-----------------------------------------------------------------------------------------------#
# Reading the allele frequency into the script. The file can be anywhere, just copy the
path of the file here so that it can be found.
open ( ALLELEFREQ, “path/428Accessions.AlleleFrequencies.csv" ) || die;
        #If the file cannot be opened the script exits.

my @alleles;                            #Create an array

while ( my $line = <ALLELEFREQ> ) {             #Reading the allele frequency
        chomp $line;

        if ( $line =~ m/SNPID/ ) {      #Skip very first line
        } else {
                my @array = split ",", $line;           #Splitting each line
                my @allelfreq;
                $allelfreq[0] = $array[0];              #SNPID
                $allelfreq[1] = $array[3];              #Col allele
                $allelfreq[2] = $array[4];              #Col allele frequency
                $allelfreq[3] = $array[5];              #Non-Col allele
                $allelfreq[4] = $array[6];              #Non-Col allele frequency

                push @alleles, \@allelfreq;             #Creating an array of the arrays to
                                                        save until later in the script.
        }
}
```

```
close (ALLELFREQ);                    #Closes the allele frequency file.

#-------------------------------------------------------------------------------------------#
PART B: reading the creating the significant SNP files. This script opens each .ps file
created from EMMAX and pulls out the significant SNPs for each phenotype. The
significant SNPs are then printed into a new file named phenotype.sigsnps.csv. There is a
file for each phenotype. If  nothing is printed then there were no SNPs that passed the
threshold of significance. The threshold is determined by the user.
#-------------------------------------------------------------------------------------------#
my $dir_list = `ls`;
my @folders = split "\n", $dir_list;

system ( "mkdir SignificantSNPs" );  #makes a folder called SignificantSNPs

#The following loop is going to go through the array @folders, which contains the names
of the folders. It will exclude any of the folders/files specified in the "elsif" conditions
because they do not contain .ps files and kill the script if not excluded. Depending on
your directory you can delete lines or add as many conditions as you need to only include
the EMMAX result folders.
for ( my $h = 0; $h <= $#folders; $h++ ) {
        my @manhan_array;
        if ( $folders[$h] =~ m/ManhanFiles/ ) {                    #Excluding
        } elsif ( $folders[$h] =~ m/manhattanfiles/ ) {            #Excluding
        } elsif ( $folders[$h] =~ m/Significant/ ) {               #Excluding
        } elsif ( $folders[$h] =~ m/428Accessions/ ) {             #Excluding
        } elsif ( $folders[$h] =~ m/Calculate/ ) {                 #Excluding
        } elsif ( $folders[$h] =~ m/Finding/ ) {                   #Excluding
        } elsif ( $folders[$h] =~ m/Leaf/ ) {                      #Excluding
        } elsif ( $folders[$h] =~ m/Stem/ ) {              #Excluding
        } else {                                        #Will open .ps files of EMMAX
output files
                my $folder = "$folders[$h]";
                my $file = "$folders[$h].ps";
                print "$folder/$file\n";              #Prints which file it is opening.
                open ( PVALUES, "$folders[$h]/$folders[$h].ps" ) || die;
        #-------------------------------------------------------------------------------#
        The following is the script that splices the .ps file into different elements and
        saves them into an array so that I can use the data later in this script.
        #-------------------------------------------------------------------------------#
        my @significantsnps; my $total = 0;
                while ( my $line = <PVALUES> ) {          #Reads the .ps files
                        chomp $line;

                        my @array = split "\t", $line;
```

```
                #Splitting the SNPID into different parts to get the
            chromosome and SNP position
            my @chrom = split "", $array[0];
                my $bp = substr( $array[0], 1 ); # ",", $array[0], 2; #SNP bp
                #Compares SNPID of .ps file to the SNPID of allele frequency
                array. If they match then the script continues.
                if ( $array[0] == $alleles[$total][0] ) {
                    #Specify your significance cut-off here
                    if ( $array[2] <= 1e-5 ) {
                        my @sigsnp;
                        $sigsnp[0] = $array[0];         #snpid
                        $sigsnp[1] = $chrom[0];         #chromosome
                        $sigsnp[2] = $bp;               #basepair
                        $sigsnp[3] = $array[2];         #p-values
                        $sigsnp[4] = $array[1];         #beta values
                        $sigsnp[5] = $alleles[$total][4];#allele frequency
                        #Creating an array of arrays of the significant
                        SNPs of an single phenotype to be printed later.
                        push @significantsnps, \@sigsnp;
                    }
                    $total++;
                }
            }
            close( PVALUES );                           #Closing the .ps file
 #-------------------------------------------------------------------------#
 This part makes a new file named phentoype.sigsnps.csv and prints all the
 significant SNPs with their p-value, beta-value, and allele frequency. The columns
contained in the file are the trait, the SNPID, the chromosome, the base pair, the
p-value, the beta-value, and the non-Columbia allele frequency. The files are saved
in a folder named SignificantSNPs
 #-------------------------------------------------------------------------#
            #Open the folder "SignificantSNPs" to print new files
            system ( "cd SignificantSNPs" );
            system ( "mkfile -nv 100k SignificantSNPs" );       #Makes a file
            #Creates and opens the sigsnps.csv file
            open (FILE, ">SignificantSNPs/$folders[$h].sigsnps.csv" ) || die;
      print FILE "Trait,SNP,CHR,BP,P,Beta,Non-Col_AlleleFreq\n";
            for ( my $r = 0; $r <= $#significantsnps; $r++ ) {
                #Loop to print the significant SNPs
                print FILE "$folders[$h],$significantsnps[$r][0],
                $significantsnps[$r][1],$significantsnps[$r][2],
                $significantsnps[$r][3],$significantsnps[$r][4],
                $significantsnps[$r][5]\n";
            }
            close ( FILE );                 # Closes the file
```

```
        }
    }
```

Appendix E FindingSignificantSNPs_NMA.pl

The difference between FindingSignificnatSNPs.pl and FindingSignificantSNPs_NMA.pl is in Part A of the script. The bolded script is the change.

```
#-----------------------------------------------------------------------------------------------#
PART A: Reading the allele frequencies into the script. THe first line is of the allele
frequency file is ommitted so that the first SNP in the arrays match the first SNP in the
arrays created using the .ps files. This is important for increasing speed of this scrip.
**The variation on this script is that SNPs with an allele frequency less than a specified
number is eliminated from the array. This script is used when you are using a genotype
file      that includes only the SNPs with a certain allele frequency. It is used so that a
new allele frequency file does not have to be created.***
#-----------------------------------------------------------------------------------------------#

# Reading the allele frequency into the script. The file can be anywhere, just copy the
path of the file here so that it can be found.
open ( ALLELEFREQ, "path/428Accessions.AlleleFrequencies.csv" ) || die;

my @alleles;                                    #Create an array

while ( my $line = <ALLELEFREQ> ) {             #Reading the allele frequency
        chomp $line;

        if ( $line =~ m/SNPID/ ) {              #Skip very first line
        } else {
                my @array = split ",", $line;   #Parse line
                my @allelfreq;
                $allelfreq[0] = $array[0];      #SNPID
                $allelfreq[1] = $array[3];      #Col allele
                $allelfreq[2] = $array[4];      #Col allele frequency
                $allelfreq[3] = $array[5];      #Non-Col allele
                $allelfreq[4] = $array[6];      #Non-Col allele frequency
                #Specify the allele frequency cut off, usually 5% or 10%.
                if ( $allelfreq[2] >= 0.05 && $allelfreq[2] <= 0.95 ) {
                        push @alleles, \@allelfreq;    #Creating an array of the arrays
                }
        }
}
close (ALLELFREQ);              #Closes the allele frequency file.
```

Appendix F CalculateFDR.pl

```
#! usr/bin/perl

#-----------------------------------------------------------------------------------------------#
This script is to calculate the FDR for GWAS analyses and to find the signifcant SNPs
using FDR.
#-----------------------------------------------------------------------------------------------#

use strict;

#-----------------------------------------------------------------------------------------------#
PART A: Reading the allele frequencies into the script. THe first line is of the allele
frequency file is ommitted so that the first SNP in the arrays match the first SNP in the
arrays created using the .ps files. This is important for increasing the speed of this scrip.
#-----------------------------------------------------------------------------------------------#

# Reading the allele frequency into the script. The file can be anywhere, just copy the
path of the file here so that it can be found.
open ( ALLELEFREQ, "/path/428Accessions.AlleleFrequencies.csv" ) || die;

my @alleles;

while ( my $line = <ALLELEFREQ> ) {
        chomp $line;

        if ( $line =~ m/SNPID/ ) {
        } else {
                my @array = split ",", $line;
                my @allelfreq;
                $allelfreq[0] = $array[0];              #SNPID
                $allelfreq[1] = $array[1];              #Col allele
                $allelfreq[2] = $array[2];              #Col allele frequency
                $allelfreq[3] = $array[3];              #Non-Col allele
                $allelfreq[4] = $array[4];              #Non-Col allele frequency
                push @alleles, \@allelfreq;
        }
}
close (ALLELFREQ);
#-----------------------------------------------------------------------------------------------#
#-----------------------------------------------------------------------------------------------#
PART B: Calculating FDR and finding the significant SNPs based on new threshold. The
FDR is calculated using the Benjamini-Hochberg method. The p-values are ranked, and a
new significant cutoff is calculated based on the ranking. The very first test is a simple
```

Bonferroni test [0.05/total tests]. The second test is [(0.05*2)/total tests], and so on. This script open each .ps file from EMMAX output and orders the SNPs based on p-value. It then ranks each p-value and calculates the new significant cutoff. If the SNP passes, the SNP is saved and printed in a new significant SNP file withe the ending .fdr.csv.

```
#-----------------------------------------------------------------------------------------------#
my $dir_list = `ls`;
my @folders = split "\n", $dir_list;

#Create folder called FDR_SignificantSNPs
system ( "mkdir FDR_SignificantSNPs" );

#This loop is going to go through my array that contains the names of the folders and
open each .ps file and does the file manipulations and prints the .manhan file.
for ( my $h = 0; $h <= $#folders; $h++ ) {
        my @manhan_array;
        #The following loop is going to go through the array @folders, which
        contains the names of the folders. It will exclude any of the folders/files
        specified in the "elsif" conditions because they do not contain .ps files and kill
        the script if not excluded. Depending on your directory you can delete lines
        or add as many conditions as you need to only include the EMMAX result folders.
        if ( $folders[$h] =~ m/ManhanFiles/ ) {                 #Excluding
        } else {                                        #Will open .ps files of EMMAX output files
                my $folder = "$folders[$h]";
                my $file = "$folders[$h].ps";
                print "$folder/$file\n";
                open ( PVALUES, "$folders[$h]/$folders[$h].ps" ) || die;

                my @snpid; my @SNParray; my $total = 0;

                while (my $line = <PVALUES> ) {
                        chomp $line;

                        my %SNPs;                               #Define a hash
                        my @array = split "\t", $line;
                        push @snpid, $array[0];                 #Define SNP id
                        my $snpid = $array[0];                  #Define SNP id
                        my $betavalue = $array[1];              #Define beta value
                        my $pvalue = $array[2];                 #Define p-value
                        #Putting the SNPID, beta- and p-value into a hash

                        %SNPs = ( SNPID => "$snpid",
                                        PVALUE => "$pvalue",
                                         BETA => "$betavalue" );
                        push @SNParray,\%SNPs;     #Makes an array of hashes
                        $total++;
```

```
		}
		close (PVALUES);

		system ( "cd FDR_SignificantSNPs" );
		system ( "mkfile -nv 100k FDR_SignificantSNPs" );
		open (FILE, ">FDR_SignificantSNPs/$folders[$h].fdr.csv" ) || die;
		print FILE "Trait,SNP,CHR,BP,P,Beta,Non-Col_AlleleFreq,
			FDR-derivedSignificantThreshold,
			FDR-adjustedP-value,Rank\n";
		my $order = 1;#Defining variable for rank of p-values
		my @fdr_allele;
		#The following will do the actual FDR calculations#
	foreach my $snp (sort { $a->{PVALUE} <=> $b->{PVALUE} } @SNParray) {
		#Sorting all SNPs based on the p-value
		my $notsignificant = 0;
		#Calculate the new significant cutoff
		my $sigThreshold = (0.05 * $order)/($total);
		#Calculate the new adjusted p-value based on rank
		my $adjustedPvalue = $$snp{'PVALUE'} * ($total / $order);
		#Determine if p-value is less than new significant cutoff. If yes then
		define new variables
		if ($$snp{'PVALUE'} <= $sigThreshold ) {
			my @significantSNP;
			$significantSNP[2] = $$snp{'SNPID'};	#SNPID
			$significantSNP[3] = $$snp{'PVALUE'};	#Original p-value
			$significantSNP[4] = $$snp{'BETA'};	#Beta-value
			$significantSNP[5] = $sigThreshold;#New Significant threshold
			$significantSNP[6] = $adjustedPvalue;	#Adjusted p-value
			$significantSNP[7] = $order;		#Rank
			my @little_array = split "", $significantSNP[2], 2;
			$significantSNP[0] = $little_array[0];	#Chromosome
			$significantSNP[1] = $little_array[1];	#Basepair position
			#Push array of information of new significant SNP onto
			another array
			push @fdr_allele, \@significantSNP
		} else {
			#To specify that the p-values are no longer significant

		$notsignificant = 1;
		}
		$order++;		#Increase the rank number
		#If true then script ends foreach loop and continues on with the scrip
		to print the significant SNPs
		last if ( $notsignificant == 1 );
}
```

```
#Sorts all the significant SNPs based on SNP ID
my @sorted = sort { $a->[0] <=> $b->[0] || $a->[1] <=> $b->[1] } @fdr_allele;
for ( my $g = 0; $g <= $#fdr_allele; $g++ ) {#For loop to print significant SNPs
        for ( my $f = 1; $f<=$#alleles; $f++ ) {#For loop for allele frequencies
                if ( $sorted[$g][2] == $alleles[$f][0] ) {#Matching SNP ids to SNP ids of
                                                        allele frequencies
                print FILE "$folders[$h],$sorted[$g][2],$sorted[$g][0],
                $sorted[$g][1],$sorted[$g][3],$sorted[$g][4],$alleles[$f][4],
                $sorted[$g][5],$sorted[$g][6],$sorted[$g][7]\n";
                        last;
                }
        }
}
}
}
```

Appendix G DetermineGenes_LinkedToSigSNPs.pl

```
#!/usr/bin/perl

#-------------------------------------------------------------------------------------------------#
This script uses gene files downloaded from arabidopsis.org to link genes to the
significant SNPs of determined from EMMAX or MLMM. 11 genes are matched to one
SNP. The hit gene is the that contains the SNP or is closest to the SNP and then the 5
genes upstream and downstream of the hit gene. A file is printed for each phenotype
containing all the putative genes. This script should be saved in the "SignificantSNPs" or
"FDR_SignificantSNPs" folders. A new folder is created called "CandidateGenes."

#-------------------------------------------------------------------------------------------------#

use strict;

#-------------------------------------------------------------------------------------------------#
Part A: This first part of the script parses information from the file
TAIR10_functional_descriptions downloaded from arabidopsis.org. Each line of the
description file is spliced and the data specific parts of the data are saved as different
arrays.
 #-------------------------------------------------------------------------------------------------#

my @gene_description;
my $sample = 0;

open ( GENE, "/path/TAIR10_functional_descriptions") or die;  #Open file. User will
have to change the path of the file
while ( my $line = <GENE> ) {
        chomp $line;

        my @array = split "\t", $line, 5; #Splitting each line into different elements
        my @gene_name = split "", $array[0];#Splitting gene name Ex.AT1G010104.1
        $#gene_name = $#gene_name - 2;#Deleting last two digits of gene name (.1)
        $array[0] = join "", @gene_name;                #Creating gene Ex. AT1G010104
        if ( $array[0] ne $sample ) {   #Making sure gene is not already part of list.
                push @gene_description, \@array;
                #If not then gene name will be added to array
                $sample = $array[0];
                        #Setting variable of gene name, to be tested against the next
                        gene name. This is to eliminate the same gene showing up
                multiple times
        }
```

```
}

#-----------------------------------------------------------------------------------------------#
Part B: The second part of the script parses the information in the file
TAIR10_GFF3_genes.gff. The information is put into a hash, each element is defined
with a different key. The hash is needed so that I can sort the information according to
chromosome and start_bp since the genes are not in numerical order in the file. Once
each gene is in order the elements are put into different arrays so that they will be easier
to compare and keep in order since a hash does not worry about the order of things.

#-----------------------------------------------------------------------------------------------#
my ( @chrom_list, @start_bp, @end_bp, @gene_list, @molecule_list, @encoded_gene,
@encoded_genename, @other_gene_name );
my @happy;

my $y = 1;

open ( GFF, "/path/TAIR10_GFF3_genes.gff" ) or die;  #Opening file. The path will have
to be changed to where this file is saved
while ( my $line = <GFF> ) {
        chomp $line;
        #Eliminating any mitochondrial or chloroplastic genes
        if ( $line =~ m/^ChrM/ || $line =~ m/^ChrC/ ) {
        } else {
                my @array = split "\t", $line;

                if ( $array[2] eq "gene" ) {
                        #Only looking for genes defined as "gene"
                        my %gene; my $gene_descript; my $other_gene_name;
                        my $chromosome = substr( $array[0], 3 );     #Chromosome
                        my @gene_information = split ";", $array[8];
                        my $gene = substr( $gene_information[2], 5 ); #Gene name
                        my $start_number = $y;
                        for ( $y = $start_number; $y <= $#gene_description; $y++) {
                        #Links the gene description found in the
                        TAIR10_functional_description file to the rest of the
                        information of the gene
                                if ( $gene_description[$y][0] eq $gene ) {
                                        $gene_descript = join ( ';', split ",",
                                                $gene_description[$y][2] );
                                        my @other_name = split ";",
                                        $gene_description[$y][4];
                                        $other_gene_name = $other_name[0];
```

```
                                        last if ( $gene_description[$y][0] eq $gene );
                                        #Exits the loop and moves to the next
                                        step. This helps with time and memory.
                                }
                        }

                        %gene = (       chromosome => "$chromosome",
                                start_bp => "$array[3]",
                                end_bp  => "$array[4]",
                                gene => "$gene",
                                molecule_list => "$array[2]",
                                encoded_gene => "$gene_descript",
                                other_name => "$other_gene_name"
                            );
                push @happy, \%gene;
                }
        }
}

#Sorting the hashes and then putting the elements into arrays.
foreach my $protein (sort {$a->{chromosome} cmp $b->{chromosome} || $a->{start_bp}
<=> $b->{start_bp} } @happy) {
        push @chrom_list, $$protein{chromosome};
        push @start_bp, $$protein{start_bp};
        push @end_bp, $$protein{end_bp};
        push @gene_list, $$protein{gene};
        push @molecule_list, $$protein{molecule_list};
        push @encoded_gene, $$protein{encoded_gene};
        push @other_gene_name, $$protein{other_name};
}
#-----------------------------------------------------------------------------------------------#
#-----------------------------------------------------------------------------------------------#
#  Part C: The last part of the script puts everything together. The significant file is read a
line at a time. The SNP is compared to the start and end basepair of every end. If the SNP
falls between the start and end base pair of a gene then that gene is the hit gene. Then the
five genes downstream of the hit gene are printed and the 5 genes upstream of the hit
gene are printed. If the SNP does not fall in between any gene then it compares the SNP
between the end base pair of the immediate downstream gene and the start base pair of
the immediate upstream gene. The hit gene is the gene in between the downstream and
upstream genes. Once again the five closest downstream and 5 closest upstream are
printed for each SNP.
#-----------------------------------------------------------------------------------------------#
my $dir_list = `ls`;
my @folders = split "\n", $dir_list;
```

```
system ( "mkdir CandidateGenes" );  #Creating file called "CandidateGenes"

#This loop is going to go through my array that contains the names of the files in the
        SignificantSNPs folder
for ( my $h = 0; $h <= $#folders; $h++ ) {
        my @manhan_array;
        if ( $folders[$h] =~ m/CandidateGenes/ ) {                      #Excluding
        } elsif ( $folders[$h] =~ m/Determine/ ) {                      #Excluding
        } else {
                        #Opening only files that are .sigsnps.csv or fdr.csv
                my $folder = "$folders[$h]";
                my @candiarray;
                open ( SNPS, "$folders[$h]" ) || die;  #Open .sigsnps.csv file

                while ( my $line = <SNPS> ) {         #Reading line of file
                        chomp $line;
                        my @array = split ",", $line;
                        my $trait = $array[0]; #Name of phenotype
                        my $snpid = $array[1];          #SNPID
                        my $snp_chromosome = $array[2];    #Chromosome
                        my $snp = $array[3];                    #Base pair
                        my $beta = $array[5];           #Beta-value
                        my $pvalue = $array[4];                 #p-value
                        my $allelefreq = $array[6];             #Allele frequency

                        #Loop to go through gene list and match to SNPs
                        for (my $x = 0; $x <= $#gene_list; $x++ ) {
                                my $up = $x + 1;        #Define gene position in gene list
                                my $down = $x - 1;    #Define gene position in gene list
                                my $gene_twodown = $x - 2; #etc.
                                my $gene_threedown = $x - 3;
                                my $gene_fourdown = $x - 4;
                                my $gene_fivedown = $x - 5;
                                my $gene_twoup = $x + 2;
                                my $gene_threeup = $x + 3;
                                my $gene_fourup = $x + 4;
                                my $gene_fiveup = $x + 5;
                                my @candigenes;
                                #Match chromosome numbers
                                if ( $snp_chromosome == $chrom_list[$x] ) {
                                        if ( $snp >= $start_bp[$x] && $snp <=
                                                $end_bp[$x] ) {
                                                #If SNP is within the gene base
                                                pair positions then save the
                                                putative gene list
```

```
$candigenes[0] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Hit,$molecule_list[$x],$chrom_list[$x],$start_bp
[$x],$end_bp[$x],$gene_list[$x],$other_gene_name[$x],$encoded_gene[$x],http://www.
arabidopsis.org/servlets/TairObject?name=$gene_list[$x]&type=locus\n";
$candigenes[1] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Down5,$molecule_list[$gene_fivedown],$chrom
_list[$gene_fivedown],$start_bp[$gene_fivedown],$end_bp[$gene_fivedown],$gene_list
[$gene_fivedown],$other_gene_name[$gene_fivedown],$encoded_gene[$gene_fivedown
],http://www.arabidopsis.org/servlets/TairObject?name=$gene_list[$gene_fivedown]&ty
pe=locus\n";
$candigenes[2] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Down4,$molecule_list[$gene_fourdown],$chrom
_list[$gene_fourdown],$start_bp[$gene_fourdown],$end_bp[$gene_fourdown],$gene_lis
t[$gene_fourdown],$other_gene_name[$gene_fourdown],$encoded_gene[$gene_fourdo
wn],http://www.arabidopsis.org/servlets/TairObject?name=$gene_list[$gene_fourdown]
&type=locus\n";
$candigenes[3] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Down3,$molecule_list[$gene_threedown],$chro
m_list[$gene_threedown],$start_bp[$gene_threedown],$end_bp[$gene_threedown],$gen
e_list[$gene_threedown],$other_gene_name[$gene_threedown],$encoded_gene[$gene_t
hreedown],http://www.arabidopsis.org/servlets/TairObject?name=$gene_list[$gene_three
down]&type=locus\n";
$candigenes[4] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Down2,$molecule_list[$gene_twodown],$chrom
_list[$gene_twodown],$start_bp[$gene_twodown],$end_bp[$gene_twodown],$gene_list[
$gene_twodown],$other_gene_name[$gene_twodown],$encoded_gene[$gene_twodown]
,http://www.arabidopsis.org/servlets/TairObject?name=$gene_list[$gene_twodown]&typ
e=locus\n";
$candigenes[5] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Down1,$molecule_list[$down],$chrom_list[$do
wn],$start_bp[$down],$end_bp[$down],$gene_list[$down],$other_gene_name[$down],$
encoded_gene[$down],http://www.arabidopsis.org/servlets/TairObject?name=$gene_list[
$down]&type=locus\n";
$candigenes[6] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Up1,$molecule_list[$up],$chrom_list[$up],$start
_bp[$up],$end_bp[$up],$gene_list[$up],$other_gene_name[$up],$encoded_gene[$up],ht
tp://www.arabidopsis.org/servlets/TairObject?name=$gene_list[$up]&type=locus\n";
$candigenes[7] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Up2,$molecule_list[$gene_twoup],$chrom_list[$
gene_twoup],$start_bp[$gene_twoup],$end_bp[$gene_twoup],$gene_list[$gene_twoup],
$other_gene_name[$gene_twoup],$encoded_gene[$gene_twoup],http://www.arabidopsis
.org/servlets/TairObject?name=$gene_list[$gene_twoup]&type=locus\n";
$candigenes[8] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Up3,$molecule_list[$gene_threeup],$chrom_list[
$gene_threeup],$start_bp[$gene_threeup],$end_bp[$gene_threeup],$gene_list[$gene_thr
```

```
eeup],$other_gene_name[$gene_threeup],$encoded_gene[$gene_threeup],http://www.ara
bidopsis.org/servlets/TairObject?name=$gene_list[$gene_threeup]&type=locus\n";
                                                $candigenes[9] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Up4,$molecule_list[$gene_fourup],$chrom_list[$
gene_fourup],$start_bp[$gene_fourup],$end_bp[$gene_fourup],$gene_list[$gene_fourup
],$other_gene_name[$gene_fourup],$encoded_gene[$gene_fourup],http://www.arabidop
sis.org/servlets/TairObject?name=$gene_list[$gene_fourup]&type=locus\n";
                                                $candigenes[10] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Up5,$molecule_list[$gene_fiveup],$chrom_list[$
gene_fiveup],$start_bp[$gene_fiveup],$end_bp[$gene_fiveup],$gene_list[$gene_fiveup],
$other_gene_name[$gene_fiveup],$encoded_gene[$gene_fiveup],http://www.arabidopsis
.org/servlets/TairObject?name=$gene_list[$gene_fiveup]&type=locus\n";
        push @candiarray, \@candigenes;
        $x++;
        last;
} elsif ( $snp >= $end_bp[$down] && $snp <= $start_bp[$up] ) {          #If SNP is
close to gene base pair positions then save the putative gene list
                                                $candigenes[0] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Hit,$molecule_list[$x],$chrom_list[$x],$start_bp
[$x],$end_bp[$x],$gene_list[$x],$other_gene_name[$x],$encoded_gene[$x],http://www.
arabidopsis.org/servlets/TairObject?name=$gene_list[$x]&type=locus\n";
                                                $candigenes[1] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Down5,$molecule_list[$gene_fivedown],$chrom
_list[$gene_fivedown],$start_bp[$gene_fivedown],$end_bp[$gene_fivedown],$gene_list
[$gene_fivedown],$other_gene_name[$gene_fivedown],$encoded_gene[$gene_fivedown
],http://www.arabidopsis.org/servlets/TairObject?name=$gene_list[$gene_fivedown]&ty
pe=locus\n";
                                                $candigenes[2] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Down4,$molecule_list[$gene_fourdown],$chrom
_list[$gene_fourdown],$start_bp[$gene_fourdown],$end_bp[$gene_fourdown],$gene_lis
t[$gene_fourdown],$other_gene_name[$gene_fourdown],$encoded_gene[$gene_fourdo
wn],http://www.arabidopsis.org/servlets/TairObject?name=$gene_list[$gene_fourdown]
&type=locus\n";
                                                $candigenes[3] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Down3,$molecule_list[$gene_threedown],$chro
m_list[$gene_threedown],$start_bp[$gene_threedown],$end_bp[$gene_threedown],$gen
e_list[$gene_threedown],$other_gene_name[$gene_threedown],$encoded_gene[$gene_t
hreedown],http://www.arabidopsis.org/servlets/TairObject?name=$gene_list[$gene_three
down]&type=locus\n";
                                                $candigenes[4] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Down2,$molecule_list[$gene_twodown],$chrom
_list[$gene_twodown],$start_bp[$gene_twodown],$end_bp[$gene_twodown],$gene_list[
$gene_twodown],$other_gene_name[$gene_twodown],$encoded_gene[$gene_twodown]
,http://www.arabidopsis.org/servlets/TairObject?name=$gene_list[$gene_twodown]&typ
e=locus\n";
```

```
                                        $candigenes[5] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Down1,$molecule_list[$down],$chrom_list[$do
wn],$start_bp[$down],$end_bp[$down],$gene_list[$down],$other_gene_name[$down],$
encoded_gene[$down],http://www.arabidopsis.org/servlets/TairObject?name=$gene_list[
$down]&type=locus\n";
                                        $candigenes[6] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Up1,$molecule_list[$up],$chrom_list[$up],$start
_bp[$up],$end_bp[$up],$gene_list[$up],$other_gene_name[$up],$encoded_gene[$up],ht
tp://www.arabidopsis.org/servlets/TairObject?name=$gene_list[$up]&type=locus\n";
                                        $candigenes[7] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Up2,$molecule_list[$gene_twoup],$chrom_list[$
gene_twoup],$start_bp[$gene_twoup],$end_bp[$gene_twoup],$gene_list[$gene_twoup],
$other_gene_name[$gene_twoup],$encoded_gene[$gene_twoup],http://www.arabidopsis
.org/servlets/TairObject?name=$gene_list[$gene_twoup]&type=locus\n";
                                        $candigenes[8] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Up3,$molecule_list[$gene_threeup],$chrom_list[
$gene_threeup],$start_bp[$gene_threeup],$end_bp[$gene_threeup],$gene_list[$gene_thr
eeup],$other_gene_name[$gene_threeup],$encoded_gene[$gene_threeup],http://www.ara
bidopsis.org/servlets/TairObject?name=$gene_list[$gene_threeup]&type=locus\n";
                                        $candigenes[9] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Up4,$molecule_list[$gene_fourup],$chrom_list[$
gene_fourup],$start_bp[$gene_fourup],$end_bp[$gene_fourup],$gene_list[$gene_fourup
],$other_gene_name[$gene_fourup],$encoded_gene[$gene_fourup],http://www.arabidop
sis.org/servlets/TairObject?name=$gene_list[$gene_fourup]&type=locus\n";
                                        $candigenes[10] =
"$trait,$snpid,$beta,$pvalue,$allelefreq,Up5,$molecule_list[$gene_fiveup],$chrom_list[$
gene_fiveup],$start_bp[$gene_fiveup],$end_bp[$gene_fiveup],$gene_list[$gene_fiveup],
$other_gene_name[$gene_fiveup],$encoded_gene[$gene_fiveup],http://www.arabidopsis
.org/servlets/TairObject?name=$gene_list[$gene_fiveup]&type=locus\n";
                                        push @candiarray, \@candigenes;
                                        $x++;
                                }
                        }
                }
        }
        close ( SNPS );                                 #Close file
        system ( "cd CandidateGenes" );       #Change to CandidateGenes folder
        system ( "mkfile -nv 100k CandidateGenes" );         #Create file
        #Name file phenotype.candigenes.csv
        open (FILE, ">CandidateGenes/$folders[$h].candigenes.csv");
        print FILE "Trait,SNP,B_value,
        P-value,Non_Col_Allele_freq,Position,Type,Chromosome,Start_bp,
        End_bp,Gene,Name,Description,Hyperlink\n";        #Column names
        for ( my $k = 0; $k <= $#candiarray; $k++ ) {
```

```
                        print FILE
"$candiarray[$k][0]$candiarray[$k][1]$candiarray[$k][2]$candiarray[$k][3]$candiarray[
$k][4]$candiarray[$k][5]$candiarray[$k][6]$candiarray[$k][7]$candiarray[$k][8]$candia
rray[$k][9]$candiarray[$k][10]\n";  #Print gene lists
                }
                close ( FILE );                 #Close file
        }
}#-----------------------------------------------------------------------------------------------#
```

Appendix H CreatingSignificantSNPFiles.pl

```
#!/usr/bin/perl

use strict;
#-----------------------------------------------------------------------------------------------#
#  This script is going to take the SNPs from the MLMM analyses that are saved from R
and turn  them into the regular Significant SNP files that I make for the EMMAx output.
#-----------------------------------------------------------------------------------------------#
system ( "mkdir SignificantSNPs" );
my $dir_list = `ls`;
my @folders = split "\n", $dir_list;
my ( @BIC, @BONF, @STEPS, @BICp, @BONFp, @STEPSp, @STEPSinitialp );

for ( my $t = 0; $t <= $#folders; $t++ ) {
        if ( $folders[$t] =~ m/BICcof/ ) {
                push @BIC, $folders[$t];
        } elsif ( $folders[$t] =~ m/BONFcof/ ) {
                push @BONF, $folders[$t];
        } elsif ( $folders[$t] =~ m/STEPScof/ ) {
                push @STEPS, $folders[$t];
        } elsif ( $folders[$t] =~ m/BICpvalues/ ) {
                push @BICp, $folders[$t];
        } elsif ( $folders[$t] =~ m/BONFpvalues/ ) {
                push @BONFp, $folders[$t];
        } elsif ( $folders[$t] =~ m/STEPSpvalues/ ) {
                push @STEPSp, $folders[$t];
        }  elsif ( $folders[$t] =~ m/STEPSinitialpvalues/ ) {
                push @STEPSinitialp, $folders[$t];
        }
}
#-----------------------------------------------------------------------------------------------#
for ( my $h = 0; $h <= $#STEPS; $h++ ) {
        open ( SIGSNPs, "$STEPS[$h]" ) || die;
        print "$STEPS[$h]\n";

        my @detailarray;
        while ( my $line = <SIGSNPs> ) {
                chomp $line;

                if ( $line =~ m/x/ ) {
                } else {
                        my @array = split " ", $line;
                        my @chrom = split "", $array[0];
```

```
                my @details;
                $details[1] = $chrom[0];        #chromosome
                $details[2] = $array[1];        #bp
                $details[0] = "$chrom[0]$array[1]";  #snpid
                push @detailarray, \@details;
        }
}
close (SIGSNPs);

print "$STEPSinitialp[$h]\n";
open ( IPVALUES, "$STEPSinitialp[$h]" ) || die;
while ( my $line = <IPVALUES> ) {
        chomp $line;

        if ( $line =~ m/SNP/ ) {
        } else {
                my @array = split ",", $line;  #2 elements, chrom/bp and pvalue
                my @snpinfo = split " ", $array[0];             #2 chrom, bp
                my @chrom = split "", $snpinfo[0];              #split chrom-
                my @ipvalue;
                $ipvalue[1] = $chrom[0];                #chromosome
                $ipvalue[2] = $snpinfo[1];              #bp
                $ipvalue[0] = "$chrom[0]$snpinfo[1]";       #snpid
                $ipvalue[3] = $array[1];                #initial p-value
                for ( my $x = 0; $x <= $#detailarray; $x++ ) {
                if ( $detailarray[$x][0] == $ipvalue[0] ) {
                        #To push the initial pvalue onto the array.
                        push @{ $detailarray[$x] }, "$ipvalue[3]";
                }
        }
        }
}

print "$STEPSp[$h]\n";
open (PVALUES, "$STEPSp[$h]" ) || die;
while ( my $line = <PVALUES> ) {
        chomp $line;

        my @array = split ",", $line;
        my @pvalue_info;
        $pvalue_info[1] = $array[1];                    #chromosome
        $pvalue_info[2] = $array[2];                    #bp
        $pvalue_info[0] = "$array[1]$array[2]";     #snp id
        $pvalue_info[3] = $array[3];                    #pvalue
        for ( my $x = 0; $x <= $#detailarray; $x++ ) {
```

```
                        if ( $detailarray[$x][0] == $pvalue_info[0] ) {
                                #To push the pvalue onto the array.
                                push @{ $detailarray[$x] }, "$pvalue_info[3]";
                        }
                }
        }
#-------------------------------------------------------------------------------------------------#
        system ( "cd SignificantSNPs" );
        system ( "mkfile -nv 100k SignificantSNPs" );
        open (FILE, ">SignificantSNPs/$STEPS[$h].sigsnps.csv");
        print FILE "Trait,SNP,CHR,BP,InitialP,P,Beta,Non-Col_AlleleFreq\n";
        for ( my $y = 0; $y <= $#detailarray; $y++ ){
                print FILE "$STEPS[$h],$detailarray[$y][0],$detailarray[$y][1],
                $detailarray[$y][2],$detailarray[$y][3],$detailarray[$y][4],,\n";
        }
        close ( FILE );
}
#-------------------------------------------------------------------------------------------------#
for ( my $h = 0; $h <= $#BIC; $h++ ) {
        open ( SIGSNPs, "$BIC[$h]" ) || die;
        print "$BIC[$h]\n";

        my @detailarray;
        while ( my $line = <SIGSNPs> ) {
                chomp $line;

                if ( $line =~ m/x/ ) {
                } else {
                        my @array = split " ", $line;
                        my @chrom = split "", $array[0];
                        my @details;
                        $details[1] = $chrom[0];        #chromosome
                        $details[2] = $array[1];        #bp
                        $details[0] = "$chrom[0]$array[1]"; #snpid
                        push @detailarray, \@details;
                }
        }
        close (SIGSNPs);

        print "$BICp[$h]\n";
        open (PVALUES, "$BICp[$h]" ) || die;
        while ( my $line = <PVALUES> ) {
                chomp $line;

                my @array = split ",", $line;
```

```
                my @pvalue_info;
                $pvalue_info[1] = $array[1];                    #chromosome
                $pvalue_info[2] = $array[2];                    #bp
                $pvalue_info[0] = "$array[1]$array[2]";         #snp id
                $pvalue_info[3] = $array[3];                    #pvalue
                for ( my $x = 0; $x <= $#detailarray; $x++ ) {
                        if ( $detailarray[$x][0] == $pvalue_info[0] ) {
                                #To push the pvalue onto the array.
                                push @{ $detailarray[$x] }, "$pvalue_info[3]";
                        }
                }
        }
#-------------------------------------------------------------------------------------------------#
        system ( "cd SignificantSNPs" );
        system ( "mkfile -nv 100k SignificantSNPs" );
        open (FILE, ">SignificantSNPs/$BIC[$h].sigsnps.csv");
        print FILE "Trait,SNP,CHR,BP,P,Beta,Non-Col_AlleleFreq\n";
        for ( my $y = 0; $y <= $#detailarray; $y++ ){
                print FILE "$BIC[$h],$detailarray[$y][0],$detailarray[$y][1],
                        $detailarray[$y][2],$detailarray[$y][3],,\n";
        }
        close ( FILE );
}
#-------------------------------------------------------------------------------------------------#
for ( my $h = 0; $h <= $#BONF; $h++ ) {
        open ( SIGSNPs, "$BONF[$h]" ) || die;
        print "$BONF[$h]\n";

        my @detailarray;
        while ( my $line = <SIGSNPs> ) {
                chomp $line;

                if ( $line =~ m/x/ ) {
                } else {
                        my @array = split " ", $line;
                        my @chrom = split "", $array[0];
                        my @details;
                        $details[1] = $chrom[0];        #chromosome
                        $details[2] = $array[1];        #bp
                        $details[0] = "$chrom[0]$array[1]";             #snpid
                        push @detailarray, \@details;
                }
        }
        close (SIGSNPs);
```

```
        print "$BONFp[$h]\n";
        open (PVALUES, "$BONFp[$h]" ) || die;
        while ( my $line = <PVALUES> ) {
                chomp $line;

                my @array = split ",", $line;
                my @pvalue_info;
                $pvalue_info[1] = $array[1];            #chromosome
                $pvalue_info[2] = $array[2];            #bp
                $pvalue_info[0] = "$array[1]$array[2]";     #snp id
                $pvalue_info[3] = $array[3];                #pvalue
                for ( my $x = 0; $x <= $#detailarray; $x++ ) {
                        if ( $detailarray[$x][0] == $pvalue_info[0] ) {
                                #To push the pvalue onto the array.
                                push @{ $detailarray[$x] }, "$pvalue_info[3]";
                        }
                }
        }
#-------------------------------------------------------------------------------------------------#
        system ( "cd SignificantSNPs" );
        system ( "mkfile -nv 100k SignificantSNPs" );
        open (FILE, ">SignificantSNPs/$BONF[$h].sigsnps.csv");
        print FILE "Trait,SNP,CHR,BP,P,Beta,Non-Col_AlleleFreq\n";
        for ( my $y = 0; $y <= $#detailarray; $y++ ){
                print FILE "$BONF[$h],$detailarray[$y][0],$detailarray[$y][1],
                        $detailarray[$y][2],$detailarray[$y][3],,\n";
        }
        close ( FILE );
}
```